Master's thesis
Astronomy

AGN clustering in the COSMOS field

Viitanen Akke Esa Tapio

August 28, 2017

Tutor: doc. Viola Allevato

Censors: prof. Alexis Finoguenov
doc. Viola Allevato

UNIVERSITY OF HELSINKI
Department of Physics

PL 42 (Gustaf Hällströmin katu 2 A)
00014 Helsingin yliopisto

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Physics

Tekijä — Författare — Author
Viitanen Akke Esa Tapio

Työn nimi — Arbetets titel — Title

AGN clustering in the COSMOS field

Oppiaine — Läroämne — Subject
Astronomy

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's thesis	August 28, 2017	64 pages

Tiivistelmä — Referat — Abstract

Active galactic nuclei (AGN) are compact, luminous objects found in the central region of many galaxies. In the standard paradigm, the AGN is fueled by accretion of matter into a supermassive black hole (SMBH). In fact, the properties of many galaxies and their respective SMBHs are linked, which hints at the importance of AGN as factors in galaxy formation and evolution. The bulk of the matter in the Universe is some form of dark matter, which is still poorly understood. AGN are biased tracers of the underlying dark matter distribution. By comparing the clustering of AGN with that of the dark matter, the bias may be quantified and further, the bias can be linked to a characteristic mass of the dark matter halo hosting the AGN.

The advent of high-resolution X-ray telescopes, namely Chandra and XMM-Newton, has made unprecedently large samples available for study. With detailed spectroscopic follow-up programs, the study of X-ray selected AGN clustering has received a major boost. The clustering measurements tell of the typical environments that are likely to host AGN and thus shed light on what actually triggers the AGN.

In this thesis, the clustering of ~ 600 X-ray selected AGN with $z < 2.5$ ($\overline{z} = 1.19$) in the COSMOS (Cosmic evolution survey) field surveyed with XMM-Newton (XMM-COSMOS) is studied. The full sample is split into subsamples based on the host galaxy stellar mass M_* and the ratio between the X-ray luminosity and the stellar mass L_X/M_* which is a proxy for the Eddington ratio. For the full sample the bias is $3.61^{+0.37}_{-0.40}$, which corresponds to a characteristic halo mass of $\log M_{\rm halo}/h^{-1}{\rm M}_\odot = 13.52^{+0.12}_{-0.16}$, consistent with the overall picture of X-ray selected AGN residing in massive haloes with $12.5 < \log M_{\rm halo}/h^{-1}{\rm M}_\odot < 13.5$. The low M_* and high M_* samples have biases $3.53^{+0.58}_{-0.70}$ and $4.13^{+0.85}_{-1.07}$, respectively and the data do not support a difference in the typical masses of the hosting haloes. For the L_X/M_* subsamples, there is marginal evidence that low L_X/M_* AGN ($\log M_{\rm halo}/h^{-1}{\rm M}_\odot = 13.52^{+0.22}_{-0.37}$) reside in more massive haloes than high L_X/M_* AGN ($\log M_{\rm halo}/h^{-1}{\rm M}_\odot = 13.29^{+0.28}_{-0.58}$). One possible explanation would be that the environment of the low L_X/M_* AGN reduces the amount of gas available for accretion and thus results in lower accretion rates.

Avainsanat — Nyckelord — Keywords
observational astronomy, galaxy, clustering, AGN, COSMOS, XMM-Newton, X-ray

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

1. Introduction

1.1 Scientific rationale

In the past decades it has become clear that at the center of virtually every galaxy with a spheroidal component (ellipticals and bulges), there is a supermassive black hole (SMBH) with a mass $\gtrsim 10^6 M_\odot$. It has been proposed that the SMBHs reach their high masses through one or more phases of intense accretion activity shining as AGN (short for either active galactic nucleus or active galactic nuclei). The AGN-galaxy co-evolution is motivated by the observed correlation between the mass of the black hole in the galaxy and other properties of the galaxy, such as the velocity dispersion in the bulge (figure 1.1). These correlations suggest that AGN are an important piece of the puzzle in the general picture of galaxy formation and evolution. The exact forms of this feedback, the accretion mechanism, and the triggering of AGN remain unclear. Given the importance of AGN feedback, a paramount open question is the source of the gas that fuels the growth of SMBHs and the resultant nuclear activity. Of similar importance is the open question of the triggering mechanism of the AGN.

The study of AGN clustering provides a unique way to address this complex picture, providing important and independent constraints on the BH-galaxy formation and co-evolution. The majority of matter in the Universe is in some form of dark matter (DM) and galaxies as well as their SMBHs are believed to populate the collapsed DM haloes and thus, baryonic matter, such as galaxies and stars, trace the underlying DM distribution. The most common way to quantify the clustering is the two-point correlation function (2PCF), which measures the excess probability above random of finding a pair of AGN (or galaxies) at separation r. The observed AGN 2PCF may be compared against accurate numerical

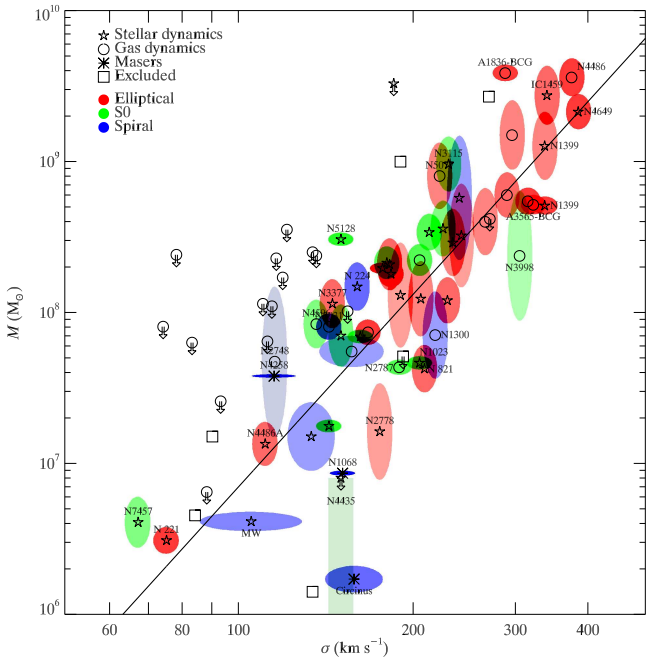

Figure 1.1: The observed correlation between the mass of the black hole M and the velocity dispersion σ in the bulge. The black line corresponds to the best-fit relation derived by Gültekin et al. (2009). Figure from Gültekin et al. (2009).

DM simulations in order to reveal the typical environments of AGN. Further, this can provide more insight about the physical mechanisms responsible for triggering AGN. The bulk of the studies done are consistent with moderate-luminosity ($L_{\rm bol} \sim 10^{44-46}$ erg s^{-1}) X-ray selected AGN residing in massive DM haloes with masses in the range of $12.5 < \log M_{\rm halo}/h^{-1}{\rm M}_\odot < 13.5$ and the triggering mechanisms of AGN is a current open problem with major galaxy mergers or disk instabilities being some of the candidate explanations (Cappelluti et al., 2012).

X-ray selection has proven to be one of the most efficient ways to sample AGN, as a luminosity of $L_X \gtrsim 10^{42}$ erg s^{-1} is almost unambigiously a sign of AGN activity. X-ray surveys are also beneficial, as the X-ray emission is not contaminated by the light of the host galaxy and thus, even low luminosity AGN may be detected. In addition, as X-ray surveys detect a large number of AGN over a large redshift range and in various cosmic environments, they are a powerful tool in order to understand BHs and galaxy evolution.

The field of X-ray astronomy is a relatively young one with first ever measurements

of the cosmic X-ray background, or CXRB, originating from the 1960s. In the following decade, *Uhuru* and *Ariel V* conducted the first all-sky X-ray surveys comprised of mostly diffuse light of unresolved objects. The next generation of observatories, such as *Einstein* and *Rosat*, saw an increase in resolution and limiting fluxes for these missions were $\sim 10^{-14}$ erg cm^{-2} s^{-1} (1−3 keV band) and $\sim 10^{-15}$ erg cm^{-2} s^{-1} (0.5−2 keV band) for *Einstein* and *Rosat*, respectively. Today, the most potent X-ray space telescopes are the two launched in 1999: *Chandra X-ray observatory* (*Chandra* for short), with subarcsec scale resolution, and the *X-ray Multi-Mirror Mission-Newton*, or *XMM-Newton*, capable of arcsec scale resolution. These high resolutions make it possible to relate the X-ray objects with their e.g. optical counterparts making multi-wavelength observations possible. *Chandra* and *XMM-Newton* observed unprecedently deep fields, as for an exposure time of $\gtrsim 75$ ks the limiting flux is 5×10^{-16} erg cm^{-2} s^{-1} in the $0.5 - 2$ keV band. As lower limiting fluxes are reached with longer exposure times, the deepest surveys are pencil-beam surveys with areas of a few deg^2 at most, tracing back the vast majority of the history of the Universe. (Brandt and Hasinger, 2005)

For comparison, in terms of detecting AGN, *XMM-Newton* and *Chandra* detected $\gtrsim 1500$ deg^{-2} AGN, while optical surveys may reach densities of $\lesssim 500$ deg^{-2}. The advent of these high-resolution X-ray telescopes, the spatial clustering measurements feature now comparable precision to large redshift surveys. The large number of sources detected by the two observatories have made it possible to study the clustering over different redshift and luminosity ranges.

1.2 Active galactic nuclei

In general, the term AGN refers to energetic phenomena, which may not be attributed to stars, at the central region of a galaxy. While stars emit light in a relatively narrow band from infrared to ultraviolet, the light from AGN is distributed all the way from the radio wavelengths to photons in the X-ray regime, featuring a non-thermal spectrum. The total energy output from AGN is also enormous and they may, in some cases, outshine the whole host galaxy.

In the AGN unified model (Urry and Padovani, 1995), the AGN is fueled by matter accretion into a SMBH, and the central region is surrounded by a dusty torus (figure 1.2). As matter is pulled toward the SMBH, a large amount of gravitational energy is released, some of which in X-rays. The compactness of the central region is highlighted by the fact that AGN may exhibit variabilities on time scales from years to months, to as short as a few days. This suggests, based on a simple causality argument, that the emitting region may be as small as a few light-days across. Material orbiting close to the BH at high velocities is photoionised, producing broad emission lines with linewidths $\sim 1000 - 20000$ km s^{-1} and is called the broad-line region (BLR). Due to the proximity of the central engine, the clouds are heated and exhibit high temperatures. The general belief among astronomers is that every AGN has a BLR, but is in some cases obscured and thus not visible in the spectrum. In some subclasses of AGN, highly collimated jets of material moving polewards close to the speed of light are found and may exit into the intergalactic space as is the case in classical radio lobes. Due to relativistic effects, the observed jets may be asymmetric as in figures 1.3 and 1.4. At a distance of up to several kpc generally also in the poleward direction is the narrow-line region (NLR). The NLR is either photoionized by the UV continuum of the central region, or excited by shocks related to the relativistic jets. The NLR lies further out, where orbital speed are also lower and the linewidths are typically of the order of $\sim 100 - 1000$ km s^{-1}. Contrary to the BLR, which may be obscured by dust, the NLR does not suffer from such obscuration and is always in view.

1.2.1 Types of AGN

Differences in AGN may be observed, for example, in their variability time scales, radio or optical properties, lumniosities, or linewidths of certain emission lines. From this wide variety of differences, many different types of AGN have been identified.

Seyfert galaxies

Named after their discoverer Carl Seyfert (1943), Seyfert galaxies are characterized by a nuclear, luminous region found typically within a late-type spiral galaxy. Seyfert galaxies are further divided into two subcategories. In their spectrum, *Seyfert I galaxies* show broad

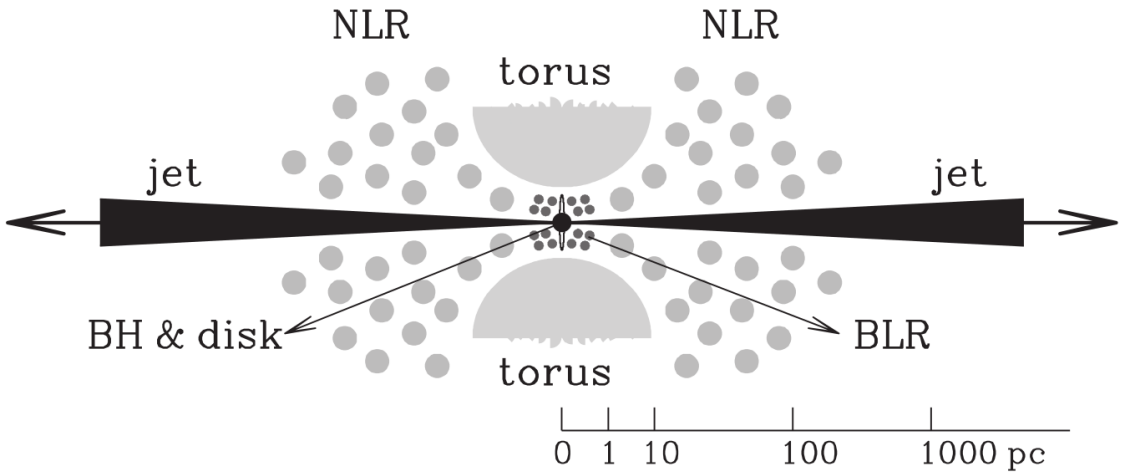

Figure 1.2: The cross-section of an AGN in the standard paradigm. At the very center of the AGN lies the SMBH, which is surrounded by the broad-line region (BLR), from which broad emission lines originate from. Based on the angle of observation, the BLR may be blocked by an obscuring torus. Further out lies the narrow-line region (NLR) that may be seen as narrow spectral lines in the spectrum. Figure from Mo et al. (2010).

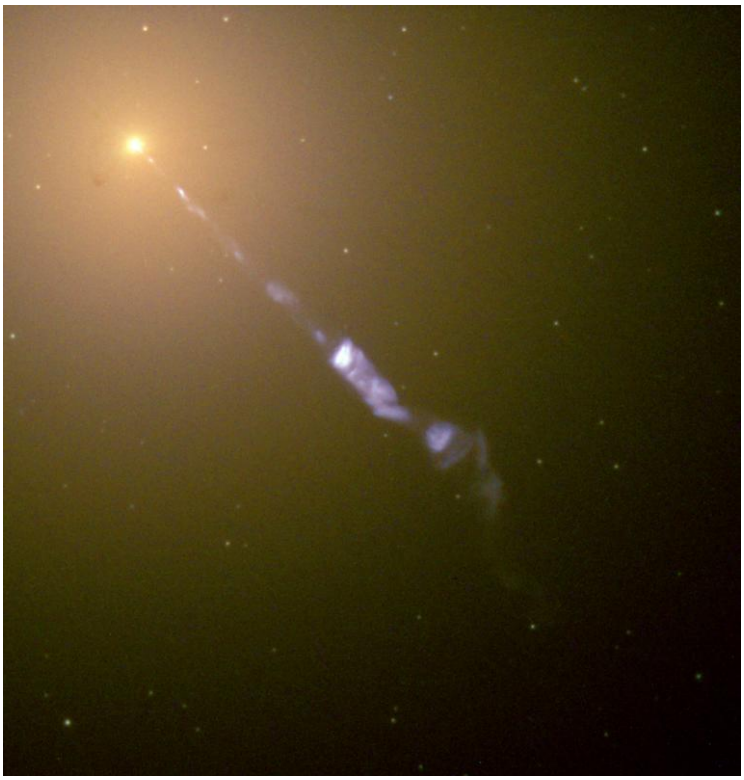

Figure 1.3: A Hubble Space Telescope (HST) optical image of the massive elliptical galaxy M87 and the spectacular jet emerging from the galaxy. Image credit: NASA and The Hubble Heritage Team (STScI/AURA) (2000).

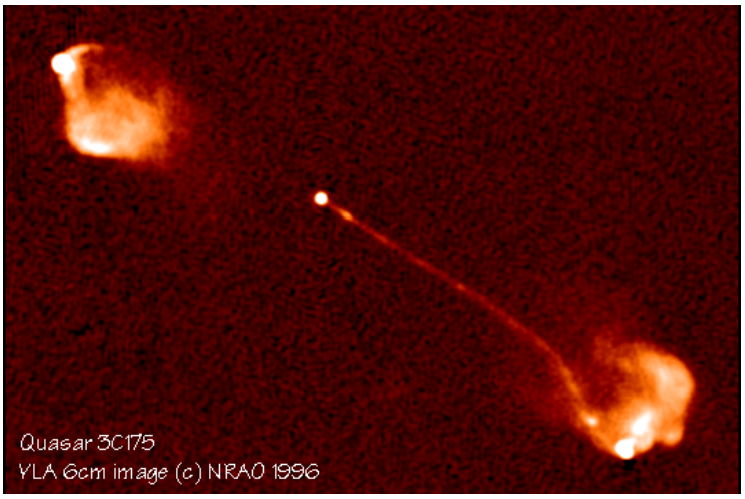

Figure 1.4: Very Large Array (VLA) image of the radio-loud quasar 3C 175 and the radio lobes. The jet appears to be one-sided, which is likely due to relativistic effects. The jet coming towards the observer is boosted in brightness, while the receding jet is dimmed and thus invisible in the image. Image credit: NRAO/AUI (1996).

permitted emission lines of e.g. hydrogen and helium with widths of ~ 10000 km s^{-1}, while the forbidden lines are approximately an order of magnitude narrower. On the contrary, the permitted and forbidden lines of *Seyfert II galaxies* have similar widths. The narrow lines originate from gas clouds at a distance of about 1 kpc from the center, while the broad lines are due to gas clouds in close proximity (~ 1 pc) to the center.

Radio galaxies

Radio galaxies are a class of objects with stronger radio emission compared to a typical spiral galaxy. Optical identifications have shown that radio galaxies are typically early-type ellipticals. In the powerful radio galaxies, two-sided radio lobes may be found that extend from hundreds of kiloparsecs to even megaparsecs from the center. Similar to Seyfert galaxies, a radio galaxy falls into one of two subcategories based on its optical spectrum: *broad-line radio galaxies* (BLRGs) or *narrow-line radio galaxies* (NLRGs). Another classification is the one introduced by Fanaroff and Riley (1974), based on radio morphology. In FR I galaxies, the distance between the two brightest spots on the opposite sides of the galaxy is less than half of the overall size of the galaxy. In FR II galaxies, this distance ratio is greater than one half.

Quasars

Quasars (*quasi-stellar radio source*) are compact radio sources that are unresolvable in the optical regime. Shortly after their discovery in the 1950s, Schmidt (1963) discovered that the quasar 3C 273 lies at an enormous redshift of $z = 2.016$. The absolute magnitude of the object in the V band is $M_V = -25.5 + 5 \log h$, which is roughly two orders of magnitude greater than that of a typical galaxy. In terms of optical spectra, quasars are similar to Seyfert I galaxies, and the distinction between the two types is that quasars are more luminous. Despite the name quasar referring to a 'radio source', these objects may be further classified as *radio-loud quasar* or *radio-quiet quasar*, depending on whether the quasar is visible in the radio. The radio-quiet quasars outnumber their radio-loud counterparts by a factor of 10 to 100.

Blazars

Blazars are a subclass of quasars and are divided into two further classes: *optically violent variables* (OVVs) and *BL Lac* objects, named after the prototype object BL Lacertae. OVVs are characterized by rapid variability in the optical on time scales of as short as days, and their spectra feature strong broad emission lines. On the contrary, the spectra of BL Lac objects are featureless power laws. Another characteristic of a blazar is the presence of *superluminal motion*, i.e. apparent velocities greater than the speed of light. Superluminal motion may be explained by particles with relativistc velocities in a narrow angle with respect to the line of sight of the observer.

Unification schemes

The idea of AGN unification is that the large variety in the observed properties of AGN may be explained with a small amount of factors. In the unified models (e.g. Antonucci 1993, Urry and Padovani 1995) the central region is obscured by a torus and the key parameter on deciding the observed properties of the AGN is the inclination angle of the torus with respect to the observer. Physically, the same object may look different based on the angle of observation.

In the *radio-quiet unification*, the differences between Seyfert I and Seyfert II galaxies are due to the obscuration of the BLR. In Seyfert I galaxies the central region is unobscured and thus broad emission lines are visible in their specta. For Seyfert II galaxies, the dusty torus obscures the central region and thus only narrow lines are visible.

The *radio-loud unification* covers radio galaxies, quasars and blazars. In NLRGs, the BLR is obscured by the torus. As the inclination of the torus grows, the observer is capable of seeing deeper in to the central region and the BLR becomes visible. With increasing inclination, the jets of the radio galaxy appear more and more luminous, in which case the AGN could be classified as a quasar. Finally, as the angle of the line of sight of the observer and the emerging jet is small, the AGN becomes a blazar.

1.2.2 Physical properties of AGN

The physical phenomena responsible for the spectrum of an AGN may be divided as follows:

- Primary – processes directly related to the black hole and the accretion disc.
- Secondary – interaction between the primary ingredients with the environment.

Examples of primary phenomena are the non-thermal spectrum and highly collimated jets, while secondary phenomena result from the former affecting the gas clouds near the black hole and the intergalactic medium (Longair, 2011).

In the vicinity of the black hole, electrons are accelerated to ultra-relativistic velocities, and spiral in the presence of a magnetic field emitting *synchrotron radiation*, largely responsible for the power-law component in the continuum of an AGN spectrum. An optically thin, physically thick accretion disc may be responsible for an ultraviolet thermal spectrum. Finally, low-energy photons scattering from the ultra-relativistic electrons may be responsible for hard X-rays and γ-rays. The process, where a relativistic electron loses energy in order to energize a photon is known as *inverse compton scattering*.

There are several key properties related to SMBHs and their accretion. A black hole is the most compact form for mass to exist and black holes are formed either by the death of massive stars, or some still relatively unknown process, e.g. a direct collapse of a giant

molecular cloud in the early Universe, which could explain the $\sim 10^9 M_\odot$ massive SMBHs already in place at $z \sim 7$ (e.g. Mortlock et al. 2011). The length scale r_g associated with a SMBH of mass M is the *Schwarzschild radius*:

$$r_g = \frac{2GM}{c^2} \simeq 3 \left(\frac{M}{M_\odot}\right) \text{ km}. \tag{1.1}$$

For a non-rotating blackhole the Schwarzschild radius also is the extent of the event horizon, and light emitted from within the event horizon suffers an infinite redshift and thus can not be observed. If the black hole possesses a finite amount of angular momentum J, it is known as a *Kerr black hole* and the surface of infinite redshift is given by:

$$r_\infty = \frac{GM}{c^2} + \left[\left(\frac{GM}{c^2}\right)^2 - \left(\frac{J}{Mc}\right)^2\right]. \tag{1.2}$$

At maximum angular momentum $J = GM^2/c$, the surface of infinity redshift occurs at $r_\infty = GM/c^2$, which is half of that in the non-rotating case. There exists a *last stable orbit* about the black hole. Within the radius of the last stable orbit, stable circular orbits do not exist, and particles on such orbits will inevitably spiral towards the black hole. For a non-rotating black hole, the last stable orbit occurs at $r = 3r_g$.

If light from the source has been observed to vary on a time scale T, the region emitting the light is causally connected. Thus, supposing that the size of the region is r_g, a lower limit for the variability time scale is given by:

$$T \geq \frac{r_g}{c} \simeq 10^{-5} \left(\frac{M}{M_\odot}\right) \text{ s}, \tag{1.3}$$

which is the causality relation.

The *Eddington luminosity* is a limit on the luminosity of a celestial body with mass M. Suppose that a proton-electron pair is at a distance of r from the light-emitting body M. The Eddington luminosity is found by equating the inward gravitational force, dominated by the mass of the proton, with the outward force felt by the electron due to the pressure of scattering photons:

$$L_{\text{Edd}} = \frac{4\pi GMm_p c}{\sigma_T} \simeq 1.3 \times 10^{38} \left(\frac{M}{M_\odot}\right) \text{ erg s}^{-1}, \tag{1.4}$$

where m_p is the mass of the proton and σ_T is the Thomson cross section. The Eddington luminosity is not a hard limit, which may not be exceeded, as a moving body or special

geometries may result in higher luminosities than the Eddington luminosity. Nevertheless, it is an extremely useful reference point for accreting bodies.

If the SMBH grows by accreting matter, its maximum luminosity may be expressed as:

$$L = \varepsilon \dot{m} c^2 \left(\frac{r_g}{r} \right), \tag{1.5}$$

where $\varepsilon \leq 1$ is the efficiency of the mass accretion and r is the radius of the object. Solving for $\dot{m}$ and including the Eddington luminosity, the mass accretion rate is:

$$\dot{m} = 1.3 \times 10^{31} \frac{L}{L_{\rm Edd}} \left(\frac{r}{r_g} \right) \frac{M}{\varepsilon M_\odot c^2}, \tag{1.6}$$

i.e. the mass of the SMBH grows exponentially $M \propto e^{t/\tau}$, with the e-folding time scale:

$$\tau \simeq 4.5 \times 10^7 \left(\frac{\varepsilon}{0.1} \right) \left(\frac{L_{\rm Edd}}{L} \right) \text{ years}. \tag{1.7}$$

This means that an initially low seed mass, could indeed grow to a massive SMBH in a time of $\lesssim 10^9$ years, but only by efficiently and constantly accreting at the Eddington limit.

1.3 Overview of this thesis

This thesis work is organized as follows:

- Chapter 2 reviews the cosmological background and discusses the large scale structure in the Universe in detail. The statistic to commonly quantify the clustering of AGN, the two-point correlation function (2PCF), is introduced and modeled based on the underlying DM distribution and the halo model.

- In chapter 3, statistical methods, such as the χ^2 minimization technique, that are frequently used in clustering analyses are discusses and algorithms that efficiently compute two-point statistics, such as counts of pairs at varying separations, are presented.

- Chapter 4 describes the XMM-Newton survey in the COSMOS field, XMM-COSMOS (~ 1.5 Ms, ~ 2 deg^2), and the XMM-COSMOS catalogs used in the clustering analysis described in the following chapter.

- Chapter 5 introduces the library used to estimate the 2PCF of XMM-COSMOS AGN and measures the clustering properties of AGN with known spectroscopic redshifts ($n_{\rm obj} = 632$) and subsamples that are based on the stellar mass and the Eddington ratio of the AGN host galaxy. The results are discussed and compared to previous studies of similar X-ray selected AGN samples and concluding remarks are given in chapter 6.

In this thesis, where it is applicable, a flat ΛCDM cosmology with $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$ and $h = 0.7$ is adopted, where $h \equiv H_0/100$ km s^{-1} Mpc^{-1}.

2. The two-point correlation function

2.1 Cosmological background

The cosmological principle states that the Universe is homogeneous and isotropic. Homogenity refers to the Universe appearing the same no matter which location one chooses and isotropy means that there are no preferred directions in the Universe. On small scales neither of these properties obviously hold as matter has formed planets and stars, which make the space inhomogeneous and anisotropic on these scales. Thus homogenity and isotropy are thought to hold in a statistical sense, meaning that given a large enough region, the Universe appears to be homogeneous and isotropic.

In a homogeneous and isotropic Universe, distances between galaxies or any arbitrary events are dictated by the Robertson-Walker metric:

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - \mathrm{d}l^2 \tag{2.1}$$

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - a(t)^2\left[\frac{\mathrm{d}r^2}{1-Kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2\right], \tag{2.2}$$

where $K = 0$, $K = +1$ and $K = -1$ correspond to a geometrically flat, spherical or hyperbolic Universe, respectively. The *scale factor* $a(t)$ allows for expansion or contraction of space in time. The *proper distance* between two objects is the integral of equation 2.2 with $\mathrm{d}t = 0 \Leftrightarrow t = \text{const}$. Consider the proper distance between $r' = 0$ and $r' = r$ with $\mathrm{d}\theta = \mathrm{d}\phi = 0$. The proper distance l is:

$$l = a(t)\int_0^r \frac{\mathrm{d}r'}{\sqrt{1-Kr'^2}} = a(t)\chi(r), \tag{2.3}$$

where $\chi(r)$ is called the *comoving distance* i.e. the proper distance in units of the scale factor. For a flat geometry, $K = 0$, the comoving distance reads simply $\chi(r) = r$. The proper distance is additive. If two objects are on a radial line separated by $\Delta l = l_2 - l_1$, it may be expressed as:

$$\Delta l = a(t) \int_0^{r_2} \frac{\mathrm{d}r'}{\sqrt{1 - Kr'^2}} - a(t) \int_0^{r_1} \frac{\mathrm{d}r'}{\sqrt{1 - Kr'^2}} \tag{2.4}$$

$$\Delta l = a(t) \int_{r_1}^{r_2} \frac{\mathrm{d}r'}{\sqrt{1 - Kr'^2}}. \tag{2.5}$$

In most cases, at cosmological scales, the distance to the object is inferred from the *redshift*, defined as

$$1 + z \equiv \frac{\lambda_o}{\lambda_e}, \tag{2.6}$$

where λ_e and λ_o are the wavelengths of a photon at the time of emission and observation, respectively. If the time between the emission and the observation is considerably large, the scale factor $a(t)$ changes in-between the two events and it is straightforward to show that

$$1 + z = \frac{\lambda_o}{\lambda_e} = \frac{a(t_o)}{a(t_e)}, \tag{2.7}$$

where the subscripts have their usual meanings. The above equation holds for objects that have no peculiar velocities, that is, objects that are said to move only with the *Hubble flow*. If the peculiar velocity of the object is significant, the distance measurement is distorted. This is the case in e.g. galaxy clusters, where objects that are physically in close proximity appear farther away from each other due to a large velocity dispersion.

The evolution of the scale factor $a(t)$ is dictated by the *Friedmann equation*[1]:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{Kc^2}{a^2} + \frac{\Lambda c^2}{3}, \tag{2.8}$$

where ρ is the density of the Universe (in units of energy per volume) and Λ is the cosmological constant introduced by Einstein in order to obtain a static Universe. The density component is typically distributed between matter ρ_m, radiation ρ_r and the cosmological constant (or vacuum energy), which can be thought of having an energy density of

[1]The Friedmann equation may be derived from general relativity, which is beyond the scope of this work. For the interested, see e.g. Mo et al. (2010).

$\rho_\Lambda = \Lambda c^2/8\pi G$. As $a(t)$ changes, matter behaves as $\rho_m \propto a(t)^{-3}$, radiation as $\rho_r \propto a(t)^{-4}$ and $\rho_\Lambda = \text{const}$ [2] so that equation 2.8 may be written as:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\left[\rho_{m,0}\left(\frac{a_0}{a}\right)^3 + \rho_{r,0}\left(\frac{a_0}{a}\right)^4 + \rho_{\Lambda,0}\right] - \frac{Kc^2}{a^2}, \tag{2.9}$$

where the subscript 0 refers to the current value of the quantity i.e. $x_0 \equiv x(t = t_0)$. The term on the left-hand side of is also known as the Hubble parameter $H(t) = \dot{a}/a$.[3] By solving equation 2.8 with respect to ρ with $K = 0$ one finds the *critical density*:

$$\rho_{\text{crit}} \equiv \frac{3H_0^2(t)}{8\pi G}, \tag{2.10}$$

and it is customary to write the densities in terms of *density parameters* defined as $\Omega \equiv \rho/\rho_{\text{crit}}$. Thus, combining equations 2.6 and 2.9:

$$H(z) = H_0\left[\Omega_{m,0}(1+z)^3 + \Omega_{r,0}(1+z)^4 + \Omega_{\Lambda,0} + \Omega_{K,0}(1+z)^2\right]^{1/2} \tag{2.11}$$

$$H(z) \equiv H_0 E(z) \tag{2.12}$$

where the curvature term has been included by defining $\Omega_{K,0} \equiv -Kc^2/H_0^2 a_0^2$.

Now, to reconcile the redshift of the object with its comoving distance, consider a ray of light which was emitted at time t and the ray of light is observed at time t_0. Rays of light obey the light-like condition $\mathrm{d}s^2 = 0$ (equation 2.2). Thus it follows that for a radial ($\mathrm{d}\phi = \mathrm{d}\theta = 0$) ray of light:

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - a(t)^2\frac{\mathrm{d}r^2}{1-Kr^2} = 0 \tag{2.13}$$

$$c\int_t^{t_0}\frac{\mathrm{d}t'}{a(t')} = \int_0^r \frac{\mathrm{d}r'}{\sqrt{1-Kr'^2}}. \tag{2.14}$$

Note that the right-hand side of the last equality is the comoving distance $\chi(r) = l/a(t)$ (equation 2.3) and thus the comoving distance is related to the scale factor via:

$$\chi(r) = c\int_t^{t_0}\frac{\mathrm{d}t'}{a(t')} = c\int_a^{a_0}\frac{\mathrm{d}a}{a\dot{a}}. \tag{2.15}$$

[2] The scalings are due to from the equation-of-state parameter w, which is different for each species. By assuming $w_\Lambda = -1$ it follows that $\rho_\Lambda = \text{const}$.

[3] Typically the Hubble parameter is given in units of km s^{-1} Mpc^{-1} with the intuition that a galaxy at a distance of one Mpc recedes at a velocity of the value of one Hubble parameter.

Further, using the redshift relation of the scale factor (equation 2.7) the comoving distance may be written as:

$$\chi(r) = \frac{c}{H_0 a_0} \int_0^z \frac{\mathrm{d}z}{E(z)}. \tag{2.16}$$

In a *flat* ΛCDM *universe* (section §2.2), where $\Omega_{m,0} + \Omega_{\Lambda,0} = 1$ and $\Omega_i = 0$ for the rest, one finds $\chi = r$ and the integral reads:

$$a_0 r = \frac{c}{H_0} \int_0^z \frac{\mathrm{d}z}{\left[\Omega_{\Lambda,0} + \Omega_{m,0}(1+z)^3\right]^{1/2}}, \tag{2.17}$$

which holds as long as $z \ll z_{\mathrm{eq}}$, where z_{eq} is the redshift of matter-radiation equality. Typically equation 2.17 is integrated numerically. Since the current value of the Hubble parameter H_0 is subject to change due to e.g. increase in measurement accuracy, it is customary to represent values using the dimensionless Hubble parameter h:

$$h \equiv \frac{H_0}{100 \text{ km s}^{-1} \text{ Mpc}^{-1}}. \tag{2.18}$$

2.2 Large scale structure

Galaxies are not distributed uniformly in the Universe, but are part of the large scale structure (LSS, figure 2.1). The LSS is made of long filamentary structures and relatively empty voids that fill the space between the filaments. Galaxy groups and clusters with sizes of $1 - 3h^{-1}$ Mpc are found at the intersections of the filaments, while relatively few galaxies are found within the voids. The theory of the formation of the LSS states that the temperature anisotropies of the cosmic microwave background (CMB) planted the seeds of structure, which then grew gravitationally to form the galaxies of the LSS that may be observed today. The details of the LSS are related to galaxy formation and evolution physics and the cosmological parameters and thus, LSS may be used to put constraints on them. (Coil, 2013)

From the temperature fluctuations it has been inferred that the Universe is flat and baryonic matter, such as gas or stars, account for $15 - 20\%$ of all the matter in the Universe. The rest is in some form of non-baryonic dark matter that only seems to feel the effect of gravitation. Dark matter (DM) is still one of the greatest unresolved questions, but currently the cold dark matter (CDM) scenario explains observations well. In the

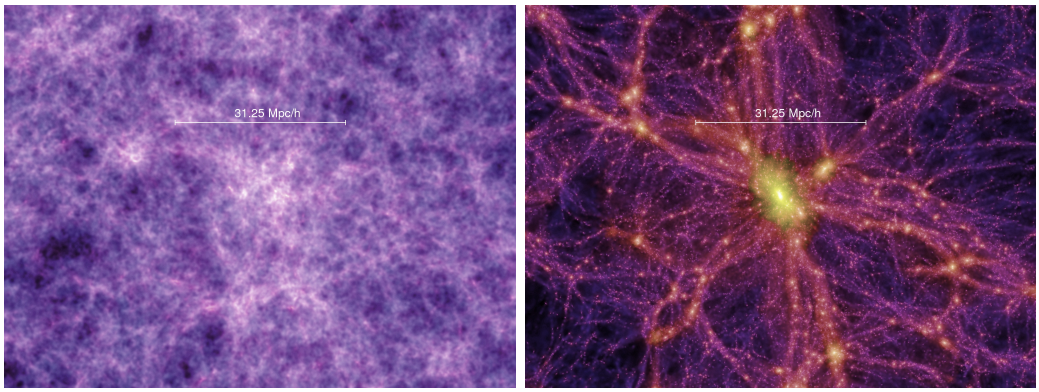

Figure 2.1: The large scale structure and its evolution in time as seen in the Millenium Simulation (Springel et al., 2005). The panels depict 15 h^{-1}Mpc slices of the distribution of dark matter at redshifts $z \sim 10$ (left-hand side) and $z = 0$ (right-hand side). The long filamentary structure may be identified as well as the voids of relatively empty space. Image credit: Springel et al. (2005).

CDM model, the DM particles were non-relativistic (cold) roughly at the time of matter-radiation equality and thus form structures more efficiently than e.g. hot or warm DM scenarios. (Roos, 2003; Mo et al., 2010)

Another mystery, even more poorly understood, is what is called dark energy. If the Universe were close to flat and dominated by matter, the expansion of the Universe would slow down significantly (flat Universe $\Omega_m = 1$) or eventually stop expanding and collapse (closed Universe $\Omega_m < 1$). In one of the most exciting discoveries, observations of distant type Ia supernovae in the late 1990s showed that the Universe is accelerating as opposed to decelerating (Riess et al., 1998; Perlmutter et al., 1999). This meant that the energy content of the Universe had to revised and the agent responsible for the accelerated expansion is commonly called *dark energy*. Very little is known about the actual nature of dark energy, for example, whether it is the cosmological constant Λ in Einstein's field equations or varies in time. An important parameter is the equation of state parameter $w = P/\rho c^2$, which relates the energy density and the pressure. The simplest equation of state parameter for a cosmological constant dark energy is $w = -1$, which corresponds to a constant energy density in time. The current cosmological model favoured by observations, i.e. the concordance model, is a flat universe with energy content dominated by dark energy and matter, a majority of which is cold dark matter. The model is commonly called flat ΛCDM, where the most recent measurements of the CMB suggest $H_0 = 67.8 \pm 0.9$ km s^{-1}

Mpc^{-1} and $\Omega_m = 0.308 \pm 0.012$ (Planck Collaboration et al., 2016).

The structure in the Universe has evolved from the nearly smooth initial distribution during the epoch of the CMB, into the complex structures observable today. The building block of the structure is a dark matter halo and as dark matter is pressureless, the growth of the formation and growth of these haloes is insensitive to complex baryonic physics. Initially the density perturbation $\delta = \delta\rho/\rho$ within a region of space grows linearly. The overdensity grows until the region is no longer expanding with the surrounding Universe, but begins to collapse. Assuming a spherical collapse model (Gunn and Gott, 1972), the structure stops collapsing as it reaches virial equilibrium, at which point the density contrast is $\delta \sim 150 - 200$, depending on the exact definition of collapse. The spherical collapse model describes the evolution of a single overdensity and a monolithic collapse. In reality the haloes grow through collapsing and merging and thus the actual evolution of the haloes is highly non-linear and may be studied with numerical N-body simulations.

Under the assumption that galaxies occupy the collapsed dark matter haloes, the clustering of galaxies may be quantified by estimating the distribution of galaxies within the haloes, i.e. what number of galaxies occupy which halo. The advantage is, that the haloes themselves are insensitive to complex baryon physics that govern the formation and evolution of galaxies and the properties of the haloes may be studied using N-body simulations or analytic calculations.

2.3 Dark matter halo properties

Dark matter haloes are roughly spherical collapsed objects. The population of different dark matter haloes is characterized by the *halo mass function*, which gives the number of haloes within a mass interval of $[M, M + \mathrm{d}M]$. Additionally, massive haloes are *biased* with respect to the underlying matter distribution and the bias is a function of the mass of the halo $b = b(M)$. Lastly, each dark matter halo has a density profile $\rho(\mathbf{r})$.

A well-known mass function for the haloes is acquired through the Press-Schechter (PS) formalism (Press and Schechter, 1974). The idea is that all of the density perturbations above some critical value $\delta_c > 1.69$ collapse to form haloes. The initial guess for the

distribution of the density perturbations is a Gaussian with zero mean and variance σ_M^2 and the probability that a given perturbation δ is between $[\delta, \delta + \mathrm{d}\delta]$ is:

$$p(\delta) = \frac{1}{\sqrt{2\pi}\sigma_M} \exp\left[-\frac{\delta^2}{2\sigma_M^2}\right]. \tag{2.19}$$

If all of the haloes with density perturbations above the critical value δ_c were to collapse and form bound objects instantly, the fraction of haloes is then given by:

$$F(M) = \int_{\delta_c}^{\infty} p(\delta) d\delta = \frac{1}{\sqrt{2\pi}\sigma_M} \int_{\delta_c}^{\infty} \exp\left[-\frac{\delta^2}{2\sigma_M^2}\right] \mathrm{d}\delta, \tag{2.20}$$

where the integral is the well-known error function:

$$\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t. \tag{2.21}$$

Thus:

$$F(M) = \frac{1}{2}\left[1 - \Phi\left(\delta/\sqrt{2}\sigma_M\right)\right], \tag{2.22}$$

where the factor $1/2$ comes from the fact that only regions satisfying $\delta > \delta_c$ collapse and regions with $\delta < -\delta_c$ do not. Eventually, the mass function at redshift z is found to be:

$$n(M, z)\mathrm{d}M = \sqrt{\frac{2}{\pi}} \frac{\overline{\rho}}{M^2} \frac{\delta_c}{\sigma_M} \exp\left(-\frac{\delta_c^2}{2\sigma_M^2}\right) \left|\frac{\mathrm{d}\ln\sigma_M}{\mathrm{d}\ln M}\right| \mathrm{d}M \tag{2.23}$$

where $\overline{\rho}$ is the mean mass density in the Universe. In figure 2.2, the mass function of DM haloes found in the Millenium Simulation (Springel et al., 2005) is compared with theoretical predictions and the PS formalism is found to agree rather well with the simulation results.

For the bias corresponding to a halo of mass M, many analytic forms exist. The typical parametrization includes the peak height parameter $\nu \equiv \delta_c/\sigma(M_{\mathrm{halo}}, z)$, where δ_c is the critical value for the collapse and $\sigma(M_{\mathrm{halo}}, z)$ is the linear density variance within spheres that enclose the mass M_{halo} at redshift z. Based on the spherical collapse model, Press and Schechter derived the bias:

$$b^{\mathrm{PS}} = 1 - \frac{\nu^2 - 1}{\delta_c}. \tag{2.24}$$

In general, the collapse does not need to happen in an axisymmetric manner and the bias parameter can be derived for an ellipsoidal collapse model. One of such is the bias derived

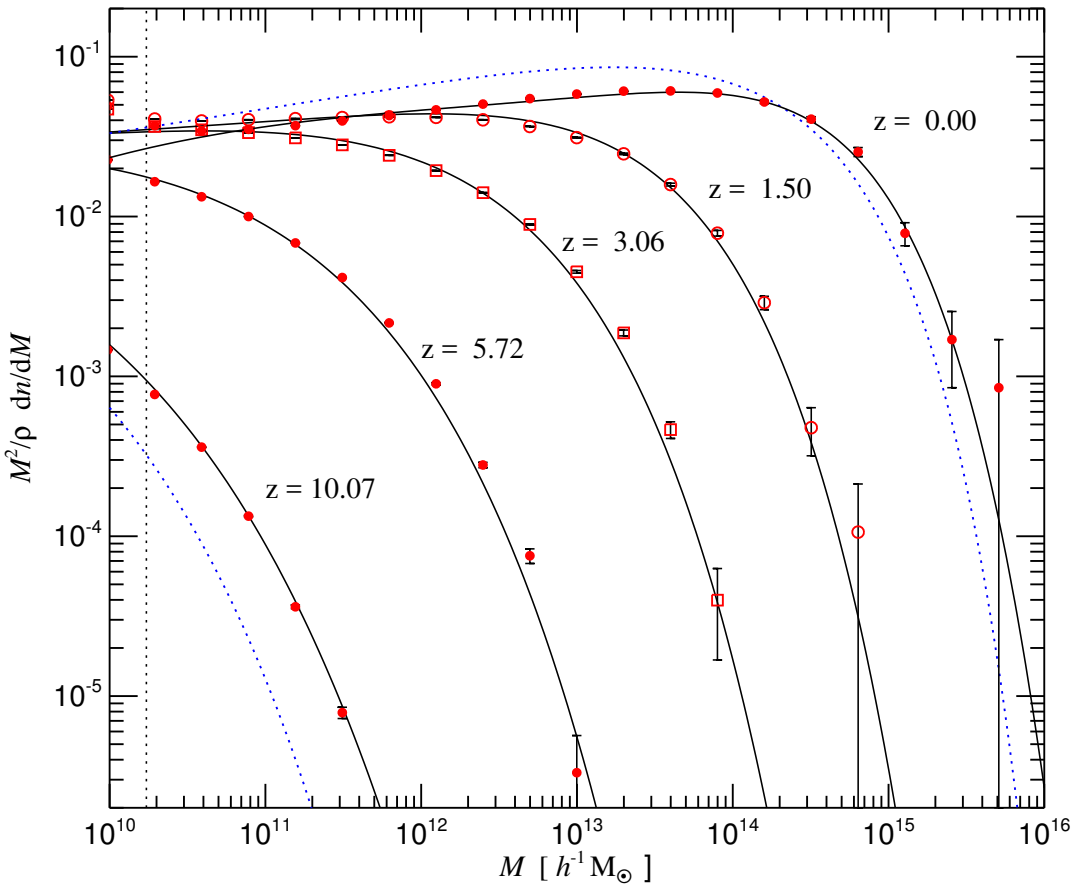

Figure 2.2: The mass function of DM haloes (red points with 1σ error bars) in the Millenium Simulation (Springel et al., 2005). The PS formalism mass function (blue dashed line) is shown for $z = 10.07$ and $z = 0$, while the black solid lines correspond to an analytic fitting function by Jenkins et al. (2001). ρ is the mean density of the Universe. Figure from Springel et al. (2005).

by Sheth and Tormen (1999):

$$b^{\mathrm{ST}} = 1 + \frac{a\nu^2 - 1}{\delta_c} + \frac{2p/\delta_c}{1 + (a\nu^2)^p}, \tag{2.25}$$

where $a = 0.707$ and $p = 0.3$ are constants. Another analytic bias based on the ellipsoidal collapse is the one provided by Sheth et al. (2001):

$$b^{\mathrm{SMT}} = 1 + \frac{1}{\sqrt{a}\delta_c}\left[\sqrt{a}(a\nu^2) + \sqrt{a}b(a\nu^2)^{1-c} - \frac{(a\nu^2)^c}{(a\nu^2)^c + b(1-c)(1-c/2)}\right], \tag{2.26}$$

where $a = 0.707$, $b = 0.5$ and $c = 0.6$. The same form of the bias, but with differently calibrated parameters yield $a = 0.707$, $b = 0.35$ and $c = 0.8$ (Tinker et al., 2005). A recipe is provided by van den Bosch (2002) for the estimation of the peak height paramater ν that has been calibrated against N-body simulations. The fitting function is applicable over a wide range of masses as well as cosmologies. In figure 2.3, the bias relation $b(M_{\mathrm{halo}}, z)$ is presented for $z = 1$ for the different bias estimates discussed. The bias relation may also be inverted in terms of the halo mass to acquire $M_{\mathrm{halo}}(b, z)$, which in turn gives the mass of the halo corresponding to a particular bias.

2.4 2PCFS

The probability of finding a galaxy at a certain point in space depends on the surroundings of the galaxy, meaning that the galaxies are clustered. The most common tool to quantify the clustering of galaxies is the two-point correlation function (2PCF) $\xi(r)$.[4] Given a galaxy, it is defined as the excess probability above random of finding another galaxy at distance r (Peebles, 1980):

$$\mathrm{d}P = n\left[1 + \xi(r)\right]\mathrm{d}V. \tag{2.27}$$

Here n is the mean number density of the galaxies and $\mathrm{d}V$ is the size of the volume element the other galaxy lies in. If all of the positions of the galaxies were to follow the Poisson distribution, it implies that $\xi(r) = 0$ at all scales and $\mathrm{d}P = n\mathrm{d}V$. Such a distribution is achieved, if each position of each galaxy were drawn uniformly inside a volume V. It is

[4] In the general case, the two-point correlation function is also a function of the location in space and the direction i.e. $\xi = \xi(\mathbf{x}+\mathbf{r})$. However, in an isotropic Universe that is homogeneous, the two-point correlation function is a function of the separation of the pair alone i.e. $\xi = \xi(r)$.

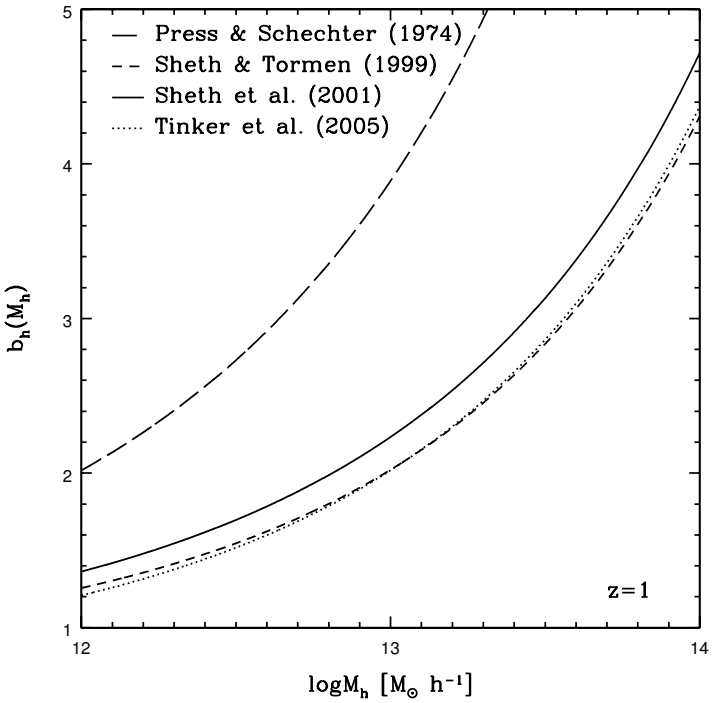

Figure 2.3: The differences between the bias estimates with a range of halo masses M_h. Image from Cappelluti et al. (2012).

also possible for $\xi(r)$ to be negative, which suggests that it is less probable to find another galaxy at the distance r compared to a Poisson distribution.

2.4.1 LS estimator

The estimation of $\xi(r)$ from a galaxy sample involves creating a random sample, which serves as an unclustered distribution of points against which the data points are compared to. The random sample and the data sample should be subject to the same observational effects. These effects might arise from geometrical constraints, i.e. the sky coverage of the data sample, the depth of the survey, and selection effects arising from e.g. inhomogeneous sky coverage of the survey. The redshift distribution of the random sample should also follow that of the data sample. However for the random sample the distribution may be smoothed in order to remove the effects of clustering present in it.

Although there are other estimators for the 2PCF used in the literature, one of the

most commonly used is the Landy & Szalay estimator, which is written symbolically as:

$$\xi_{\rm LS}(r) = \frac{[DD] - 2[DR] + [RR]}{[RR]}. \tag{2.28}$$

Here

$$[DD] \equiv DD(r)\frac{n_r(n_r-1)}{n_d(n_d-1)} \tag{2.29}$$

$$[DR] \equiv DR(r)\frac{(n_r-1)}{2n_d} \tag{2.30}$$

$$[RR] \equiv RR(r) \tag{2.31}$$

with n_d and n_r being the number of objects in the data sample and the random sample respectively. n_r should be high enough in order to not introduce Poisson errors in the estimator due to random sampling, but also as low as possible to keep computations efficiently fast. Typically values of the order of $n_r \sim 100n_d$ are adopted. $DD(r)$ (data-data), $DR(r)$ (data-random) and $RR(r)$ (random-random) are counts of pairs at separation r in the data sample and the random sample. The DR pairs include one object selected from the data sample and another from the random sample, while the other pair counts have both of the objects selected from the same sample. Typically the distances are binned so that all of the pairs at a separation of $r \pm \Delta r/2$, where Δr is the width of the bin, are included in the same bin.

In the classic Landy & Szalay estimator, each pair increments the pair count by one. However, observations of galaxies may be associated with weights which may describe e.g. the quality of the observation. In this case, the weight of the pair is $w = w_1 \times w_2$, where w_1 and w_2 are the weights of the two galaxies, and a pair with weight w increments the pair count by w, which might not necessarily be unity. Also the number counts should reflect the weights of the objects as well with $n = \sum_i w_i$, objects with smaller weights contribute less to the normalization. An immediate inspection reveals that the generalized case reduces to the classic estimator, if $w_i = 1$ for all i. The generalization is readily applied to galaxies for which a secure redshift is not available, but only a probability density function for the redshift $p(z = z_i)$ is available. In this picture, one galaxy is thought to reside at different redshifts with different weights. The pairs this one galaxy forms will then reside at different separations and with different weights. Studies using these photometric redshift probability

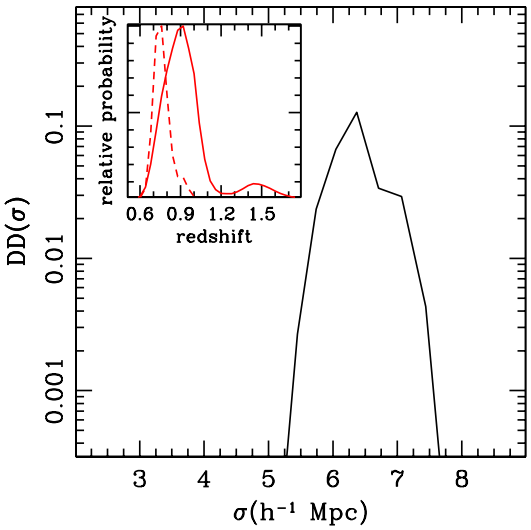

Figure 2.4: Demonstration of a single pair count using objects with redshift probability density functions. As can be seen from the inset plot, the redshifts of the two objects (the red dashed line and the red solid curve) span from $z \sim 0.6$ to $z \sim 1.6$. The pair counts from these two objects are depicted in the main plot with σ (r_p) being the perpendicular separation of the pair. The DD counts can be thought of as the convolution of the two probability density functions. Figure from Georgakakis et al. (2014).

density function have been conducted by e.g. Georgakakis et al. (2014) and Allevato et al. (2016). The downside is that the number of objects with photometric redshifts needed to recover the quality of the correlation signal achieved with spectroscopic redshifts is an order of magnitude larger (Georgakakis et al., 2014). However, spectroscopy is costly in terms of observation time compared to photometry. For example, the Euclid satellite (to be launched in 2020) will measure the spectra of several million galaxies, while photometry will be measured for more than a billion galaxies (Laureijs et al., 2011). See figure 2.4 for an illustration of using photometric redshift probability density functions.

2.4.2 Projected 2PCF

At cosmological scales, the distance from us to the object is typically inferred from the redshift of the object. This means that the distance to the object is not only dependant on the cosmological model, but also affected by the peculiar motion of the object. On small scales, such as that of galaxy groups and clusters, the galaxies exhibit large random velocities. These kinds of distortions are commonly called 'Fingers of God', as, due to

these distortions the positions of the galaxies appear elongated along the line-of-sight as if pointing at the observer. On larger scales the redshift is affected by coherent infall of galaxies into collapsing structures known as the Kaiser effect (Kaiser, 1987), which is that, to the outside observer, the coherent motions of the collapsing structure appear to contract the structure i.e. make it more compact along the line-of-sight. The two-point correlation function that has been measured using distances that are inferred from redshifts (and hence exhibit the aforementioned distortions) is called the redshift-space correlation function $\xi(s)$, where s is the redshift-space scale of interest.

As introduced by Davis and Peebles (1983), in order to get rid of the redshift-space distortions, the separation of the pair, s, may be split into two components: one parallel to the line-of-sight, π, and the other perpendicular to the line-of-sight, r_p. If $\mathbf{s} = \mathbf{s_1} - \mathbf{s_2}$ is the redshift-space separation of the pair and $\mathbf{l} = \frac{1}{2}(r_1 + r_2)$ is the mean distance to the pair, then π and r_p are defined as:

$$\pi = \frac{\mathbf{s} \cdot \mathbf{l}}{|\mathbf{l}|} \tag{2.32}$$

$$r_p = \sqrt{\mathbf{s} \cdot \mathbf{s} - \pi^2}. \tag{2.33}$$

The *projected* two-point correlation function is:

$$w_p(r_p) = \int_{-\infty}^{\infty} \xi(r_p, \pi) \mathrm{d}\pi. \tag{2.34}$$

In figure 2.5, the choice of the coordinates r_p and π is illustrated while figure 2.6 depicts the effect of redshift-space distortions on the 2PCF.

Since the line-of-sight distance is affected the most by redshift-space distortions, the projected correlation function is less senstivie to the distortions than the two-point correlation function. Further, in an statistically isotropic Universe, the two-point correlation function is symmetric, i.e. $\xi(r) = \xi(-r)$, and the projected correlation function reads:

$$w_p(r_p) = 2 \int_{0}^{\infty} \xi(r_p, \pi) \mathrm{d}\pi. \tag{2.35}$$

Ordinarily the integration is not carried to infinity, but to some limiting value of $\pi_{\max}$. The choice of $\pi_{\max}$ is a compromise between too low a value i.e. not including most of the correlated pairs thus not maximizing the correlation signal – and too high a value which

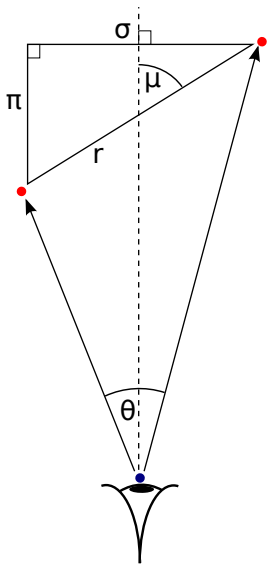

Figure 2.5: Illustration of the choice of the coordinates r_p and π. The separation between two objects r is split into one component perpendicular to the line-of-sight (dashed line), r_p (σ in the image), and another parallel to the line-of-sight, π. Figure from Alonso (2012).

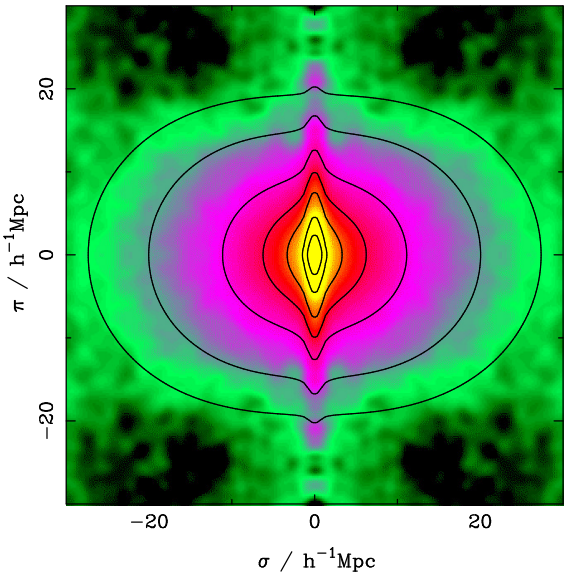

Figure 2.6: The two-point correlation function in the 2dF survey. The upper right quadrant of the image is mirrored to the other quadrants in order to reveal deviations from circular symmetry. Along the line-of-sight direction (π), the contours are narrowly elongated on small σ (r_p) scales ('fingers of God'). On large scale, the contours are flattened due to the Kaiser effect. Figure from Peacock et al. (2001).

introduces noise in the signal due to uncorrelated pairs. The $\pi_{\max}$ value is determined from the data by increasing $\pi_{\max}$ incrementally and inspecting whether the correlation signal has increased significantly or converged. Typically, values of $\pi_{\max} \sim 40h^{-1} - 80h^{-1}$ Mpc are sufficient for convergence (Coil, 2013). The integral relation in equation 2.35 may also be inverted. With a change of variable $r \equiv \sqrt{r_p^2 + \pi^2}$ it follows that $\mathrm{d}\pi = r\mathrm{d}r/\sqrt{r^2 - r_p^2}$ and equation 2.35 becomes:

$$w_p(r_p) = \int_{r_p}^{\infty} \xi(r) \frac{r\mathrm{d}r}{\sqrt{r^2 - r_p^2}}. \tag{2.36}$$

The solution for $\xi(r)$ is known as the Abel integral (Davis and Peebles, 1983):

$$\xi(r) = -\frac{1}{\pi} \int_r^{\infty} \frac{w_p'(r_p)\mathrm{d}r_p}{\sqrt{r_p^2 - r^2}}, \tag{2.37}$$

where the prime denotes differentiation with respect to the argument.

The two-point correlation function may also be defined via the *density perturbation* field. If $\rho(\mathbf{r})$ is the density (in units of mass per volume or energy per volume) at position $\mathbf{r}$, then the density perturbation is defined as:

$$\delta(\mathbf{r}) = \frac{\rho(\mathbf{r}) - \overline{\rho}}{\overline{\rho}}, \tag{2.38}$$

where $\overline{\rho}$ is the average density of the background field taken over the volume. The two-point correlation function is then:

$$\xi(r) = \langle \delta(\mathbf{r}_1)\delta(\mathbf{r}_2) \rangle, \tag{2.39}$$

where $r = |\mathbf{r}_1 - \mathbf{r}_2|$. The Fourier transform of the two-point correlation function is called the *power spectrum* $P(k)$[5]:

$$\xi(r) = \frac{1}{(2\pi)^3} \int P(k) e^{i\mathbf{k}\cdot\mathbf{x}} \mathrm{d}^3\mathbf{k}, \tag{2.40}$$

which is found by using the Fourier transform of δ and the power spectrum is related to δ via $P(k) \equiv V\langle\, |\delta_{\mathbf{k}}|\, \rangle$, where V is the volume. Since the density perturbation is dimensionless, the power spectrum has dimensions of volume and a dimensionless quantity:

$$\Delta^2(k) \equiv \frac{k^3}{2\pi^2} P(k) \tag{2.41}$$

[5]The notation of Mo et al. (2010) is followed here.

might be used instead. Note that the normalization is done in three dimensions; k has units of the inverse of length and k^3 thus has the units of inverse of volume. In spherical coordinates, one may perform the integration over the angles in equation 2.40 to obtain:

$$\xi(r) = \int_0^\infty \Delta^2(k) \frac{\sin kr}{kr} \frac{\mathrm{d}k}{k}. \tag{2.42}$$

Equation 2.40 may also be inverted for $P(k)$:

$$P(k) = 4\pi \int_0^\infty \xi(r) \frac{\sin kr}{kr} r^2 \mathrm{d}r \tag{2.43}$$

which gives that the two-point correlation function and its power spectrum are a *Fourier transform pair*.

2.5 AGN bias

The two-point correlation function is usually modeled with a power-law

$$\xi(r) = \left(\frac{r}{r_0}\right)^{-\gamma}, \tag{2.44}$$

where r_0 is a characteristic scale, where $\xi(r = r_0) = 1$ independent of γ, which gives the slope of the function. r_0 and γ may not be the same for different populations of galaxies, as different galaxies cluster in a different manner, which makes it interesting to study, what are the properties of the galaxies that affect these parameters and which properties do not. In order to gain insight on the order of r_0 and γ for X-ray selected AGN, Cappelluti et al. (2012) have gathered the results from studies done within the last 20 years, where r_0 is found to be within $5h^{-1} - 15h^{-1}$ Mpc and γ within $1 - 2$ in the majority of these studies.

Another approach is the halo model. In this framework, all of the mass in the Universe is thought to reside in distinct units (haloes) and the clustering signal is due to 1) clustering of objects within the same halo, called the 1-halo term and 2) clustering of two different haloes, called the 2-halo term. On scales less that the size of a typical halo (e.g. $\leq 1\ h^{-1}$Mpc), the 1-halo term is the dominant term in the correlation signal, while on larger scales, the 2-halo term is the significant contributor. For equation 2.35, the halo

model approach reads simply:

$$w_p(r_p) = w_p^{\text{1-h}}(r_p) + w_p^{\text{2-h}}(r_p). \tag{2.45}$$

In order to probe the clustering within the halo, *the halo occupation distribution function* needs to be modeled as well. The halo occupation distribution describes how the galaxies populate individual haloes while the halo occupation distribution function describes the probability of a halo with given mass to host a galaxy. The clustering within a halo is in the non-linear regime and thus out of the grasp of linear theory for perturbations. On the contrary, clustering on the large scale, the 2-halo term, is in the regime of linear theory. (Cooray and Sheth, 2002; Cappelluti et al., 2012)

The distribution of galaxies in the universe might not reflect that of the bulk of the matter in the Universe. For example, clusters of galaxies are rare objects that formed from the highest density peaks of the mass distribution and it is therefore possible that these objects are *biased* tracers of the underlying mass distribution.

A simple assumption for the functional form of the large scale AGN bias on the is to assume a linear bias:

$$b = \frac{\delta_{\text{AGN}}}{\delta}, \tag{2.46}$$

where δ_{AGN} and δ are the density contrasts of AGN and the background, respectively. On the other hand, the linear bias may be defined in terms of the two-point correlation function:

$$b^2 = \left(\frac{\xi_{\text{AGN}}}{\xi_{\text{DM}}}\right). \tag{2.47}$$

Solving with respect to ξ_{AGN} and integrating with respect to the line-of-sight direction π (equation 2.32) yields:

$$w_{p,\text{AGN}}(r_p) = b^2 w_{p,\text{DM}}(r_p). \tag{2.48}$$

3. Statistical methods and computational algorithms

3.1 The χ^2 minimization technique

The χ^2 distribution with k degrees of freedom is defined as:

$$\chi^2 = \sum_{i=1}^{k} X_i^2, \tag{3.1}$$

where X_i are drawn from a standard normal distribution.

Consider a scenario where there is observational data in k bins and there is a model that explains the observations which depends on a set of parameters $\mathbf{b}$. In the case that the observations are uncorrelated, the χ^2 value is

$$\chi^2 = \sum_{i=1}^{k} \frac{[O_i - f(O_i, \mathbf{b})]^2}{\sigma_i^2}, \tag{3.2}$$

where $f(O_i, \mathbf{b})$ is the model function and σ_i is the uncertainty of O_i. In general, different observations as well as their errors might be correlated and the correlations are described by the covariance matrix C, which is a $k \times k$ symmetric matrix. The elements of the covariance matrix are given by

$$C_{i,j} = \langle \left(O_i - \langle O_i \rangle\right) \left(O_j - \langle O_j \rangle\right) \rangle, \tag{3.3}$$

where $\langle X \rangle$ is the expectation value of X. Now, the χ^2 value to be minimized is given by:

$$\chi^2 = \Delta^T C^{-1} \Delta, \tag{3.4}$$

where $\Delta_i \equiv O_i - f(O_i, \mathbf{b})$ is the difference of the observation and the model and C^{-1} is the inverse of the covariance matrix. For uncorrelated errors, the covariance matrix is

diagonal:

$$C_{i,j} = \begin{cases} \sigma_i^2, & \text{for } i = j \\ 0, & \text{for } i \neq j. \end{cases} \tag{3.5}$$

Further, the inverse of the covariance matrix is simply:

$$C_{i,j}^{-1} = \begin{cases} 1/\sigma_i^2, & \text{for } i = j \\ 0, & \text{for } i \neq j \end{cases} \tag{3.6}$$

from which one can see that expressions equation 3.2 and equation 3.4 are equal for uncorrelated errors.

In the χ^2 minimization technique, one finds the set of parameters $\mathbf{b}$ which minimize the χ^2 value in equation 3.4. The uncertainty of $\mathbf{b}$ is a region in the parameter space, bounded by some limiting χ^2 value $\Delta\chi^2$. Within this region:

$$\chi^2 \leq \chi^2_{\text{min}} + \Delta\chi^2, \tag{3.7}$$

where χ^2_{min} is the minimum χ^2 value. The limiting value $\Delta\chi^2$ depends on the model (through the number of parameters) and the required level of confidence. For example, for a model with a single parameter, i.e. $\mathbf{b} = b$, then the 68 per cent confidence region for b is given by $\chi^2_{\text{min}} + 1.0$ (Avni, 1976; Wall and Jenkins, 2012).

A criteria for an appropriate model is that χ^2_{min} is approximately the same as the degrees of freedom. In detail, the degrees of freedom ν for a model with N parameters and observations in k bins is $\nu = N - k - 1$ and a criteria for a proper model is $\chi^2_{\text{min}}/\nu \sim 1$. (Wall and Jenkins, 2012)

3.2 Resampling error estimation

A first order approximation of the accuracy of the 2PCF is the Poisson error of each bin that is proportional to the square root of the pair counts in that bin. Poisson errors typically underestimate the uncertainties, as the underlying assumption is that different bins are uncorrelated. However, in general different bins are correlated with each other, since a single point – via varying distance pairs – may appear in multiple different bins. Thus accurate error estimates require knowledge of the $n \times n$ covariance matrix of the n

bins. Accuracies of particular bins may be then reported as the square root of the diagonal of the covariance matrix.

The covariance matrix may be constructed using resampling methods which are based on constructing subsamples from the available data. Two commonly used resampling methods are called the *bootstrap* method and the *jackknife* method. With the 2PCF, both of these resampling methods are based on dividing the whole survey volume into smaller subvolumes, which are then resampled in order to create new 'mock' datasets.

In the bootstrap method, one samples the subvolumes with replacement i.e. one region may appear in the new dataset more than once. Weights are then added to the regions depending on how many times the region was selected. If there are a total of N_r trials for the resampling, the covariance matrix is given by:

$$C_{\mathrm{boot}}(x_i, x_j) = \frac{1}{N_r - 1} \sum_{k=1}^{N_r} (x_i^k - \overline{x}_i)(x_j^k - \overline{x}_j), \tag{3.8}$$

where $\overline{x} = \sum_k = 1_r^N x_i^k / N_r$ is the mean of the estimated value based on the N_r trials.

In the jackknife method a new dataset is created by leaving out one of the subvolumes from the analysis. While in the bootstrap method, even with a modest number of subvolumes, the possible number of different combinations is numerous, in the jackknife, the number of different samples is the same as the number of subvolumes N_{sub}. The resampling may thus be done by leaving each of the subvolumes out one at a time. With a total number of $N_r = N$ trials, the covariance matrix is:

$$C_{\mathrm{jk}}(x_i, x_j) = \frac{N_r - 1}{N_r} \sum_{k=1}^{N_r} (x_i^k - \overline{x}_i)(x_j^k - \overline{x}_j), \tag{3.9}$$

where $\overline{x}$ is evaluated in a similar manner as in equation 3.8. (Norberg et al., 2009)

3.3 The inverse CDF method

The inverse cumulative distribution function (CDF) method is used to sample random numbers that follow some distribution that is known. If $U \in [0, 1[$ is a uniform random number. Then, it can be shown, that $F^{-1}(U)$ follows the probability distribution function $f(x)$, where $F(x)$ is the cumulative distribution function of $f(x)$. This method is valid in theory for all probability distributions $f(x)$, but the problem is that the inverse of the CDF

might not exist in closed form for all distributions. As $F^{-1}(x)$ may also be numerically approximated, the inverse CDF method is a very general and widely used for random number generation.

As an academic exercise, consider that one wants to sample random numbers from a properly normalized distribution $f(x) \propto \cos x$, where $x \in [0, \pi/2]$. The CDF is then given by

$$F(x) = \int_0^x f(x')\mathrm{d}x' \propto \sin x, \tag{3.10}$$

and the inverse of the CDF is:

$$F^{-1}(x) \propto \arcsin x \tag{3.11}$$

Thus, if $U \in [0, 1[$ is a uniform random number, then $F^{-1}(U)$ is distributed according to the wanted distribution $f(x)$. See figure 3.1 for an illustration.

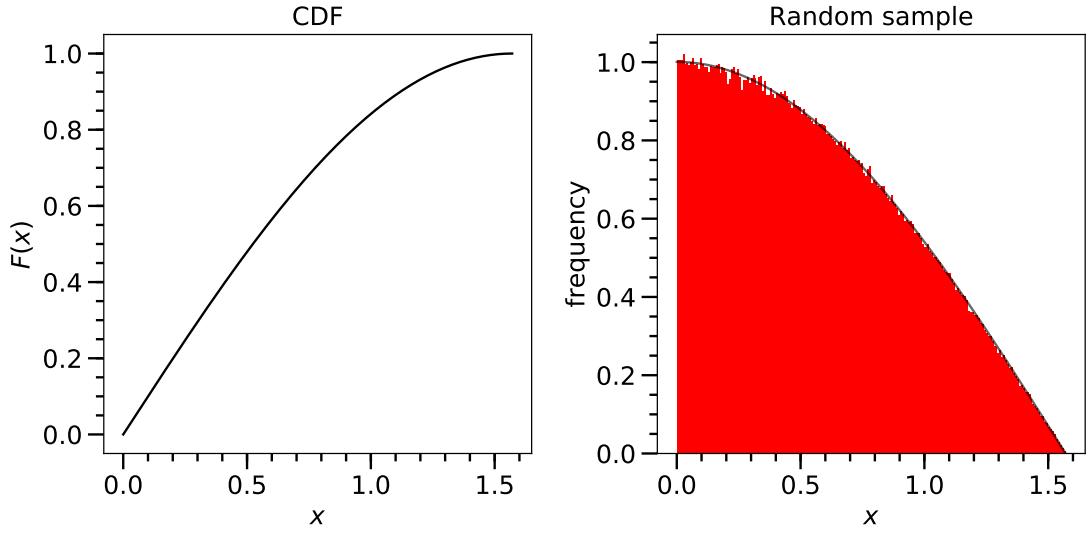

Figure 3.1: Illustration of the inverse CDF method for random number sampling. The left-hand panel shows the CDF for a probability distribution function $f(x) \propto \cos x$ and the right-hand panel shows the distribution of 10^6 random numbers $F^{-1}(U)$, where $U \in [0, 1]$ is a uniform random number.

3.4 Algorithms for efficient pair counting

To count pairs of galaxies, or any objects, is a computationally expensive task. In a total of n galaxies, for each galaxy there exists $n - 1$ pairs, which makes the total number of

pairs $n(n-1)$. If only unique pairs matter, i.e $(n_i, n_j) = (n_j, n_i)$, then total number of pairs is $n(n-1)/2$. Still, the problem of counting all of the unique pairs is on $\mathcal{O}(n^2)$ problem featuring trigonometric functions, which are expensive to calculate, to find out the separation of the pair. The total number of galaxies in the data catalog may be in the millions and the total number of objects in the random catalog is typically several orders of magnitude larger. The remedy comes from the fact that typical scales of interest are smaller than the maximum separation of any two objects. The positions of two galaxies residing gigaparsecs away from each other are likely not correlated. This means that not every pair needs to be taken into account, only the ones that are within the scales of interest. Two algorithms based on this idea of only counting the pairs that matter are represented below dubbed *ChainMesh* and *kdtree*.

3.4.1 ChainMesh

The rough idea of the ChainMesh algorithm is that the survey area is divided into smaller cells (e.g. cubical cells in a three-dimensional case) and then each galaxy is assigned a cell. Consider that the survey area is enclosed by a box parametrized by two points on the opposite sides of the box $(X_{\min}, Y_{\min}, Z_{\min})$ and $(X_{\max}, Y_{\max}, Z_{\max})$. The total number of cells in each dimension (n_x, n_y, n_z) may be decided upon based on the data or some a priori information. Now, in the case of regular, cubical cells, the cell has the dimensions of:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} (X_{\max} - X_{\min})/n_x \\ (Y_{\max} - Y_{\min})/n_y \\ (Z_{\max} - Z_{\min})/n_z \end{pmatrix}. \tag{3.12}$$

Now the maximum distance of interest may be directly related to the cells. There exists a cell that is distant enough, that any cell beyond that may not contain pairs that are closer. This means that for each galaxy that belongs in some cell, one only needs to calculate the pairs that reside near this cell. See figure 3.2 for an illustration.

The efficiency of the ChainMesh algorithm is based on only taking into account the pairs that matter, but the problem still may scale as $\mathcal{O}(n^2)$. Consider that with for each cell one needs to search for pairs in k neighbouring cells, which is some fraction of the total

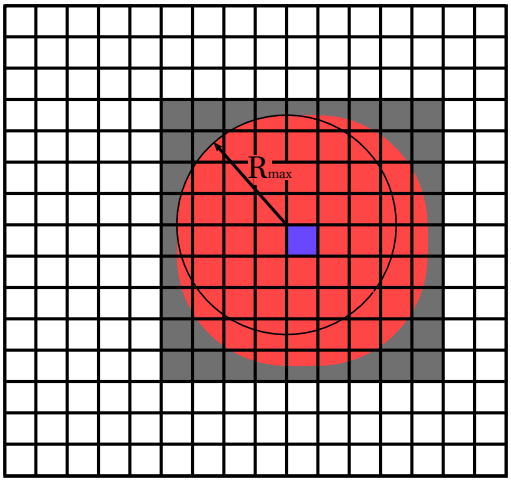

Figure 3.2: The ChainMesh algorithm saves computational time by counting only the pairs whose separation is less than the maximum scale of interest The whole survey area is divided into an evenly spaces grid of cells, or a mesh. Upon finding pairs for objects that reside in the blue cell, one only needs to search from within the gray cells, as they safely contain all of the other objects that are within $R_{\max}$. The red area corresponds to the total area that is within distance $R_{\max}$ from the blue cell. Figure from Alonso (2012).

cells $\eta \equiv k/n_{\text{cell}}$. Then, for each cell one needs to perform $\mathcal{O}(\eta(n-1)) = \mathcal{O}(n)$ calculations and the total number of calculations is still $\mathcal{O}(n^2)$. However, the creation of the mesh, i.e. assigning each object to a cell, is only $\mathcal{O}(n)$ as each object only needs to be assigned once.

3.4.2 kdtree

The word kdtree is short for k-dimensional tree. In this algorithm, the data structure used for the pair counting is a tree structure made of a *root node*, *child nodes* and *leaf nodes*, which have no child nodes. In a tree structure, each parent node may have a constant number n child nodes, where the simplest non-trivial kdtree would have $n = 2$ i.e. a *binary tree*. Further, in a tree structure, every child node has strictly one parent node.

The term k-dimensional refers to the number of dimensions in the data. When dealing with coordinates in 3-dimensional space, the choice of $k = 3$ is appropriate. In this case, each node in the tree represents a volume in the data. The root node would represent a volume that contains *all* of the points in the data. Then, the child nodes of

the parent node would effectively split the volume in two, so that the combined number of points contained in the child nodes is the same as was in the parent node. This split is based on one of the three dimensions e.g. the x coordinate of the points. A suitable choice for the split criterion could then be the mean of the x coordinate of the points. Any subsequent splits would be based on a different coordinate. For example, if d is the current level of the tree, then $d \mod k$ could correspond to the dimension of the next split, where mod is the modulo operator. Alternatively, one may split the space based on the widest dimension in the current node. See figure 3.3 for an illustration.

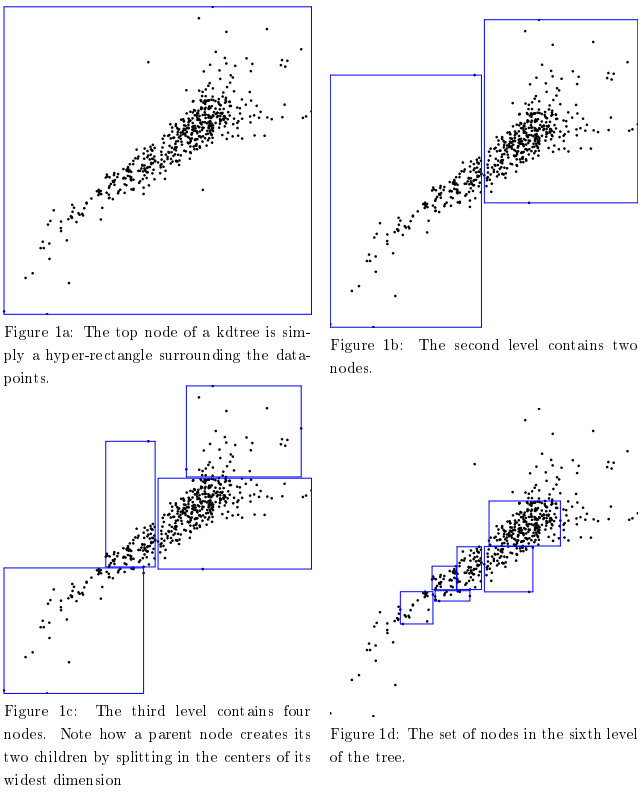

Figure 1a: The top node of a kdtree is simply a hyper-rectangle surrounding the data-points.

Figure 1b: The second level contains two nodes.

Figure 1c: The third level contains four nodes. Note how a parent node creates its two children by splitting in the centers of its widest dimension

Figure 1d: The set of nodes in the sixth level of the tree.

Figure 3.3: Depiction of the structure of a binary kdtree. The first node of the tree contains all of the points. The second level of the tree contains two nodes and the split has been performed in the vertical direction. In this case, the *parent node* is the one in the upper-left corner of the figure and the two child nodes are the nodes depicted in the upper-right. See also the captions in the image for further details. Figure from Moore et al. (2001).

After the tree is constructed, each node in the tree represents a volume in space. In the cartesian case, every node corresponds to a cubical volume parametrized by $(x_{\min}, y_{\min}, z_{\min})$

and $(x_{\max}, y_{\max}, z_{\max})$, where the minimum and maximum coordinates originate from the points residing in that node. Now, it is straightforward to calculate the minimum and maximum *distance* between any two nodes and what is more, if the minimum distance between node i and node j is greater than the maximum distance of interest, it is not possible to find relevant pairs in these two nodes. Also, since the child nodes of node i and node j may not be closer to each other than the parents, one does not need to search for pairs from the child nodes of node i and node j, thus allowing one to avoid counting uninteresting pairs. See figure figure 3.4 for an illustration.

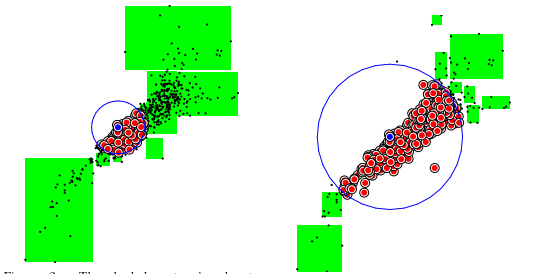

Figure 2a: The shaded rectangles denote nodes that were pruned during a search for the set of points that lie inside the circle.

Figure 2b: When the range is larger there are fewer opportunities for pruning.

Figure 3.4: Illustration how the kdtree structure is used to avoid unnecessary pair counting. Upon finding pairs for the object of interest (blue point), one does not need to search the shaded green regions as they contain points that do not lie within the blue circle, which, in turn, represents the maximum distance of interest. Depending on the maximum distance of interest, most of the points may be excluded (see the difference between the left-hand image and the right-hand image). Figure from Moore et al. (2001).

While it is clear that the kdtree algorithm is faster than a bruteforce approach, the actual performance increase is based on many factors such as the maximum distance of interest. In the preliminary tests conducted by Moore et al. (2001), they found the scaling to be $\mathcal{O}(n\sqrt{n})$ which is slightly worse than the sought-after $\mathcal{O}(n \log n)$ scalability. Nevertheless, it is preferable over the bruteforce approach as the construction of the k-d tree does not add significant overhead compared to the pair counting.

4. XMM-COSMOS

The Cosmic Evolution Survey[1] (COSMOS) is a recent survey, which features a wide and deep field covered by a multitude of different instruments at varying wavelengths. The multi-wavelength coverage extends from the radio with wavelength of tens of centimeters, all the way to the X-ray with photon energies in the keV range. A science goal of the survey is study the properties and clustering of AGN up to $z \sim 6$, which corresponds to a lookback time of ~ 10 Gyr (see figure 4.1). The center of the COSMOS field is located at RA $10^{\mathrm{h}}00^{\mathrm{m}}26.6^{\mathrm{s}}$, DEC $+02°12'21.0''$ (J2000) and covers an area of 1.4×1.4 degrees (Scoville et al., 2007), or an area of approximately ~ 2 deg^2. Since the field is located close to the equatorial, it is accessible to most ground-based instruments such as the VLA located in North America or ALMA located in South America. (Scoville et al., 2007)

XMM-Newton is an ESA space X-ray observatory which has observed the whole COSMOS field (hereafter XMM-COSMOS) for a total of ~ 1.5 Ms during the observing cycles AO3, AO4 and AO6 in three different bands: $0.5-2$ keV, $2-10$ keV and $5-10$ keV.[2] Along NASA's *Chandra*, *XMM-Newton* is the only telescope to have covered the field in the X-ray band. The *Chandra* COSMOS Legacy survey now covers an area of 2.2 deg^2 for a total exposure time of ~ 4.6 Ms (Civano et al., 2016), extending from the previous ~ 1.8 Ms and 0.5 deg^2 *Chandra* COSMOS survey (Elvis et al., 2009). Given the subarcsec resolution of *Chandra*, the present dataset will be of the highest quality to date. However, the more mature *XMM-Newton* dataset has been studied more extensively as of yet.

[1] http://cosmos.astro.caltech.edu/

[2] XMM-COSMOS was undertaken with *XMM-Newton* EPIC instrument in the bands $0.5-2$, $2-4.5$ and $4.5-10$ keV and the observations were converted into the aforementioned bands which are commonly used in the literature (Hasinger et al., 2007)

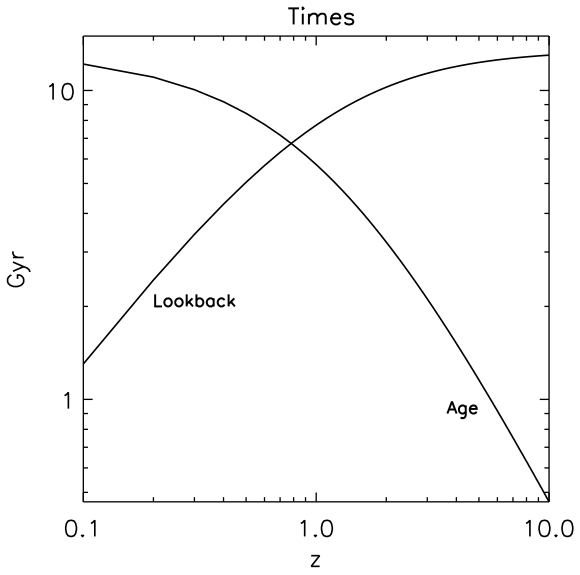

Figure 4.1: Lookback times and the age of the Universe at the relevant redshifts in the COSMOS field. Figure: Scoville et al. (2007).

Due to the observing strategy used in XMM-COSMOS, the coverage of the field is fairly uniform, but not exactly so, rather, it features stripes and regions with differing exposure times. The most abundant objects in XMM-COSMOS are AGN, followed by galaxy clusters that are extended objects and also a population stars with coronal emission may be found in the field. The XMM-COSMOS false color image of the sources and the map of the limiting fluxes are presented in figures 4.2 and 4.3, respectively. (Hasinger et al., 2007; Cappelluti et al., 2009)

4.1 XMM-COSMOS AGN

The XMM-COSMOS X-ray point-like source catalog was presented by Cappelluti et al. (2009). The limiting fluxes are 1.7×10^{-15}, 9.3×10^{-15} and 1.3×10^{-14} erg s^{-1} cm^{-2} for the $0.5-2$ keV, $2-10$ keV and $5-10$ keV bands, respectively. On another note, the faintest objects in the catalog have fluxes of 5×10^{-16}, 3×10^{-15} and 7×10^{-15} erg s^{-1} cm^{-2} for their respective bands. The catalog features a total number of 1887 sources detected in at least one of the *XMM-Newton* bands. (Cappelluti et al., 2009; Brusa et al., 2010)

The optical identifications are the key to not only connect the X-ray detections

XID	ra	dec	fluxsoft	z	w	T	$\log L_X$	$\log M_*$	err^d_m	err^u_m	$\log(Lx/M_*)$
	$\overline{\mathrm{ra}} = 150.14$	$\overline{\mathrm{dec}} = 2.25$	$\overline{f}_{\mathrm{soft}} = 6.80$	$\overline{z} = 1.30$	$\sum w_i = 1120.6$		$\log \overline{L}_X = 44.1$	$\log \overline{M}_* = 10.9$			
1	150.1051483	1.9808168	13.9	0.373	1.0	1.0	43.7	9.238	9.228	10.29	34.462
2	149.7391899	2.2205331	10.5	1.024	1.0	1.0	45.21	10.152	10.1	10.17	35.058
3	149.7615427	2.3184922	15.0	0.345	1.0	1.0	44.23	11.071	11.071	11.265	33.159
4	149.7441813	2.249476	7.84	0.132	1.0	1.0	42.56	10.598	10.597	10.598	31.962002
5	149.8281899	2.1642078	7.28	1.157	1.0	1.0	45.06	11.201	11.025	11.549	33.859
6	150.179776	2.1101542	3.82	0.36	1.0	1.0	43.45	10.975	10.906	11.067	32.475
7	150.5210769	2.625247	9.66	0.519	1.0	1.0	44.2	11.526	11.375	11.526	32.674
8	150.0538298	2.5896702	5.78	0.699	1.0	1.0	44.33	11.041	9.922	11.161	33.289
9	149.9198279	2.3274745	2.83	1.459	1.0	1.0	44.78	11.517	11.258	11.517	33.263
32	149.8592209	2.2581436	1.09	0.92	1.380608E-5	2.0	43.78	10.321	10.617	10.321	33.459
32	149.8592209	2.2581436	1.09	0.93	7.193766E-5	2.0	43.78	10.321	10.617	10.321	33.459
32	149.8592209	2.2581436	1.09	0.94	2.835534E-4	2.0	43.78	10.321	10.617	10.321	33.459
32	149.8592209	2.2581436	1.09	0.95	7.499365E-4	2.0	43.78	10.321	10.617	10.321	33.459
32	149.8592209	2.2581436	1.09	0.96	0.001703476	2.0	43.78	10.321	10.617	10.321	33.459
32	149.8592209	2.2581436	1.09	0.97	0.004149647	2.0	43.78	10.321	10.617	10.321	33.459
32	149.8592209	2.2581436	1.09	0.98	0.008964499	2.0	43.78	10.321	10.617	10.321	33.459

Table 4.1: An excerpt from the XMM-COSMOS AGN catalog. Col. 1: Unique identifier, col. 2–3: right ascension and declination (J2000) in degrees, col. 4: flux in the $0.5-2$ keV band in 10^{-14} erg s^{-1}cm^{-2}, col. 5: redshift, col 6: weight; $w = 1.0$ for objects with specz and $w < 1$ for photz objects. Note how an object with a probability density function for the redshift is presented, i.e. the same object appears at different redshifts with different weights. Col. 7: type of the object; 1.0 for unobscured AGN, 2.0 for obscured AGN. Col. 8: Luminosity in the soft band in units of erg s^{-1}, cols. $9-11$: stellar mass (and the lower and upper limits for it) of the host galaxy in units of $\mathrm{M}_\odot$, col. 12: the difference between columns 8 and 9; a proxy for the eddington ratio λ_{edd}. An object with only photometric redshifts available is presented as multiple otherwise identical objects, but with different redshifts and weights.

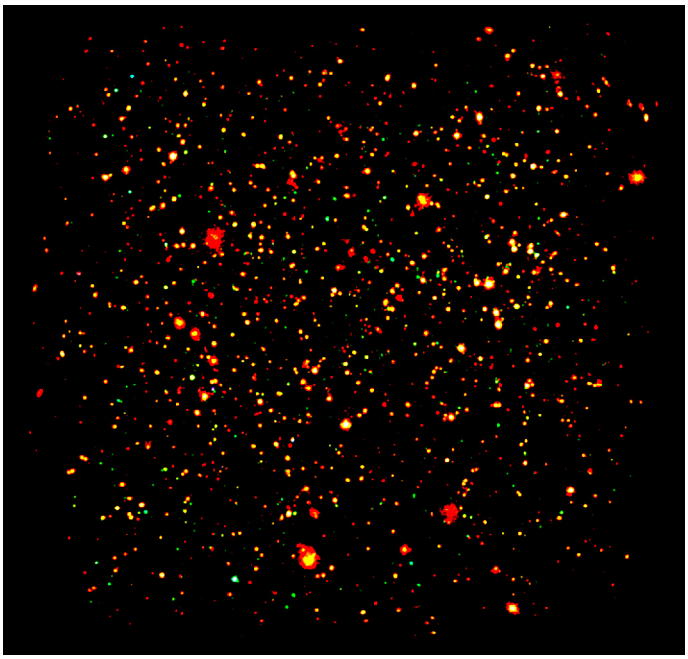

Figure 4.2: XMM-COSMOS with the colours red, green, and blue corresponding to the 0.5-2 keV, 2-4.5 keV and 4.5-10 keV bands respectively. The most abundant objects in the field are AGN while the extended objects are galaxy clusters. Figure from Cappelluti et al. (2009).

with the wide range of multiwavelength data available in the COSMOS, but also enhance the spatial resolution. *XMM-Newton* is unable to reach sub-arcsec resolution, which is attainable with *Chandra* (see e.g. figure 1 in Brusa et al. 2010). Brusa et al. (2007, 2010) reported the optical identifications for the X-ray sources in the catalog by Cappelluti et al. (2009), excluding 65 faint sources. 98 per cent of the objects have an optical counterpart. Another 95 per cent have an IRAC counterpart in the infrared and 78 per cent were also detected in 24 μm with the MIPS.[3]

The spectroscopic completeness i.e. the fraction of objects which have an observed spectrum available reported by Brusa et al. (2010) is 50 per cent, 60 per cent and 80 per cent in the $0.5-2$ keV, $2-10$ keV and $5-10$ keV bands, respectively. For these objects also the redshift is securely obtained from the spectra. For the rest, high-quality photometric redshifts are available (Salvato et al., 2009, 2011), increasing the redshift completeness to almost 100 per cent.

[3]Both IRAC (The Infrared Array Camera) and MIPS (The Multiband Imaging Photometer) are instruments aboard the Spitzer space telescope.

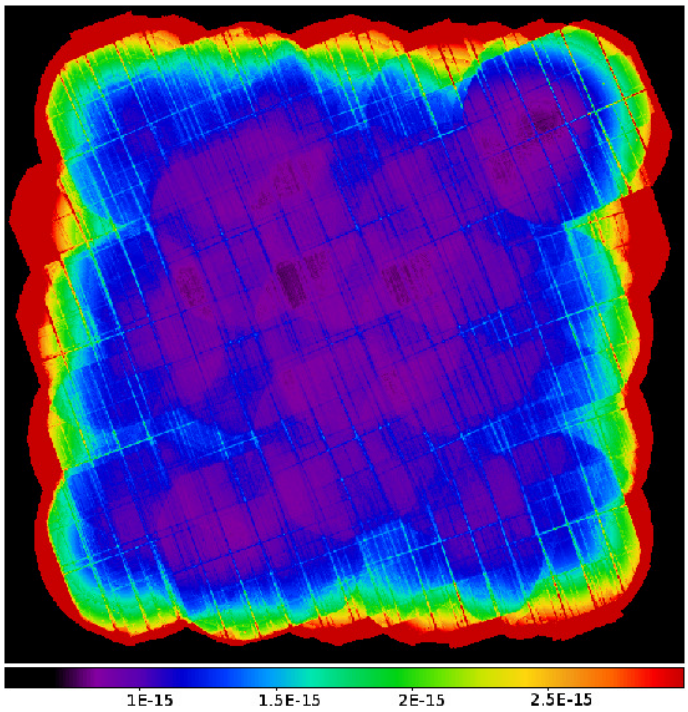

Figure 4.3: Limiting fluxes in the soft $0.5 - 2$ keV band, in units of erg s^{-1} cm^{-2}. The overall shape of the map and the non-uniformities are due to the pointing strategy of the *XMM-Newton*. Areas with lower exposure times correspond to higher limiting fluxes. Figure from Cappelluti et al. (2009)

4.2 Photometric redshifts

As opposed to spectroscopic redshifts, photometric redshifts rely on only photometric data in order to determine the redshift of the object, where the redshift acts also as a distance proxy for the object. The determination is based on explaining the SED of the object – acquired with as many low-resolution bands (e.g. optical bandpasses U,B,V,R,i) as possible – with a template spectrum such as a hybrid spectrum of a host galaxy and an AGN.

Salvato et al. (2011) extended on the work of Salvato et al. (2009) on the photometric redshift determination for the XMM-COSMOS objects. They estimated the accuracy of the redshifts by measuring the estimated photometric redshifts to the high amount of spectra available for the objects. The accuracy of the photometric redshifts is $\sigma_{\Delta z/(1+z)} \sim 0.015$ with an outlier fraction of 5.8 per cent (Salvato et al., 2009, 2011). The error reported by Salvato et al. (2011) is given in σ_{NMAD}, where NMAD stands for normalized mean absolute deviation and is defined as:

$$\sigma_{\mathrm{NMAD}} \equiv 1.48 \times \mathrm{median}\left[|z_{\mathrm{phot}} - z_{\mathrm{spec}}|/(1+z)\right]. \tag{4.1}$$

Here, z_{phot} is the probability distribution of the photometric redshift while z_{spec} is the accurate redshift as derived from emission and absorption features in the spectrum. This highlights the fact that photometric redshifts may be given by a probability distribution, rather than a single value (say, the maximum of the probability distribution) and the shape of the probability distribution is not necessarily a gaussian. For a gaussian distribution, however, σ_{NMAD} and $\sigma_{\Delta z/(1+z)}$ are directly comparable (Salvato et al., 2011). Examples of XMM-COSMOS objects with photometric redshifts are shown in fig. 4.4.

4.3 AGN host galaxy properties

Bongiorno et al. (2012) studied $\sim$ 1700 AGN in the COSMOS field – with selection based on both X-ray and optical – in order to derive the properties of the galaxies hosting the AGN. From these $\sim$ 1700 AGN, a total of 602 were classified as type I AGN (unobscured, broad lines) and the rest 1100 as type II AGN (obscured, no broad lines). Bongiorno et al. (2012) modeled the SED with a hybrid galaxy+AGN template, where both of the objects'

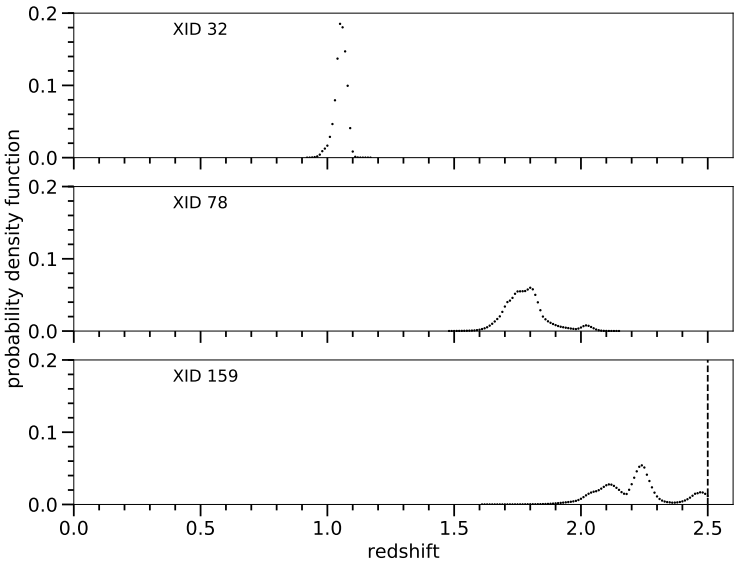

Figure 4.4: The probability density function of redshift for three selected objects. XID 32 features a nearly gaussian shape with a peak near $z \sim 1.05$. XID 78 has a more complex shape with another small maximum at $z \sim 2.05$. Finally, XID 159 shows the effect of the cut-off at $z = 2.5$ as the high-redshift tail is effectively discarded.

spectra were affected by dust reddening (see figure 4.5).

As the light of the host galaxy is largely due to the stars residing in the galaxy, Bongiorno et al. (2012) modeled the stars with a universal IMF (the same for every galaxy) and constructed differing star formation rates (SFR), each exponentially declining with the age of the galaxy t_{age} and the e-folding time τ as free parameters i.e. SFR $\propto e^{t_{\mathrm{age}}/\tau}$. Through the SED fitting Bongiorno et al. (2012) estimated the total stellar masses for 1650 AGN host galaxies. The stellar masses vary primarily from 10^{10} $\mathrm{M}_{\odot}$ to $10^{11.5}$ $\mathrm{M}_{\odot}$ with a peak at $10^{10.9}$ $\mathrm{M}_{\odot}$ (Bongiorno et al. 2012, figure 6) and there is no significant difference between the stellar masses of the hosts of type I and type II AGN. Lower stellar masses of $10^{8.5}$ $\mathrm{M}_{\odot}$ are found to be associated with optically selected Seyfert 2 galaxies with no X-ray detection.

Another quantity related to the stellar mass is the specific accretion rate, which is the ratio of the X-ray luminosity and the stellar mass L_X/M_*. This gives a fairly rough estimate of the Eddington ratio λ_{edd} (Bongiorno et al., 2012), which is defined as

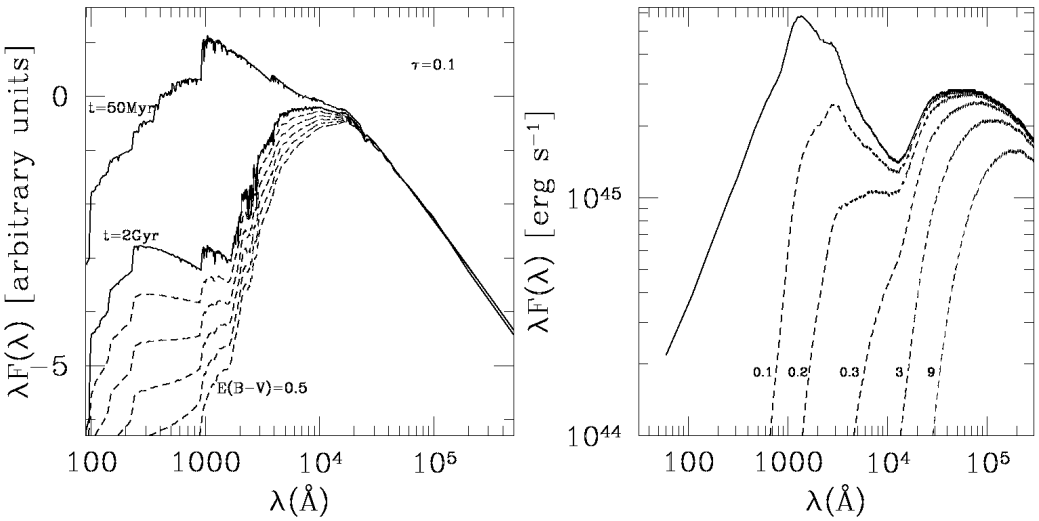

Figure 4.5: Templates for the SED fitting. Image: Bongiorno et al. (2012).

$\lambda_{\rm edd} \equiv L_{\rm bol}/L_{\rm Edd}$, where $L_{\rm bol}$ is the bolometric luminosity and $L_{\rm Edd}$ is the Eddington luminosity (Hickox et al., 2009). Explicitly, the eddington ratio may be given as:

$$\lambda_{\rm edd} = \frac{\mathcal{A} \cdot k_{\rm bol}}{10^{38}} \times \frac{L_X}{M_*}, \tag{4.2}$$

where $\mathcal{A} \sim 500 - 1000$ is a proportionality constant, which links the host stellar mass to the mass of the black hole under the assumption that the two are linked by a scaling relation. The $k_{\rm bol}$ is the $2-10$ keV bolometric correction which maps the X-ray luminosity in that band to the bolometric luminosity. Bongiorno et al. (2012) note that the $2-10$ keV bolometric correction might actually be a function of the Eddington ratio $\lambda_{\rm edd}$ itself. However, in a simple case, using a mean bolometric correction $k_{\rm bol} = 25$ and $\mathcal{A} = 500$, a ratio of $L_X/M_* \sim 10^{34}$ erg s^{-1} M$_\odot^{-1}$ corresponds approximately to an object at the Eddington limit i.e. $\lambda_{\rm edd} \sim 1$. (Bongiorno et al., 2012)

5. AGN clustering analysis

In the subsequent parts the study of the clustering properties of XMM-COSMOS AGN is presented. The projected 2PCF and the corresponding bias have been measured for the different AGN subsamples using the aforementioned methos. These methods include the creation of random catalogs and the use of full covariance matrices. A publicly available library, CosmoBolognaLib, has been used for the AGN pair counting and the error estimates.

5.1 CosmoBolognaLib

CosmoBolognaLib is an open source[1] library for cosmological calculations, going from simple distance calculations at different cosmological models to more complex measurements such as the estimation of the 2PCF. The main developer is Federico Marulli[2] and the library is currently a living project, meaning that it is actively developed. The library is written in C++, but also provides Python wrappers for simple integration in Python. Full documentation for CosmoBolognaLib is available online.[3] See also Marulli et al. (2016) for an overall description of the library and its potential.

CosmoBolognaLib only requires a galaxy catalog and a cosmological model as inputs and the library calculates the pairs and estimates the 2PCF. Although a random catalog may be created for the user automatically, the selection effects are taken better into account by constructing a random catalog manually (see section §5.2). CosmoBolognaLib also estimates the Poisson errors for the 2PCF and both the Bootstrap and Jackknife error

[1] The code is available for download at https://github.com/federicomarulli/CosmoBolognaLib

[2] federico.marulli3@unibo.it

[3] http://apps.difa.unibo.it/files/people/federico.marulli3/CosmoBolognaLib/Doc/html/

estimates have been implemented as well.

The library has the pair counting optimized using the ChainMesh algorithm described in section §3.4.1. As long as the maximum distance of interest is lower than the maximum distance between any two objects, the library will save time compared to a bruteforce approach. The ChainMesh algorithm is also parallelized using OpenMP[4], which improves the speed of the pair counting on modern multi-core machines.

5.2 Construction of the random catalog

The random catalog consists of objects that make up an unclustered distribution in the coordinate space, acting as a reference to the clustered data catalog. The random catalog must be affected by the same selection criteria and observational effects as the data catalog including, for example, the inhomogeneity of the sensitivity over the field. Following Miyaji et al. (2007) and Allevato et al. (2011):

1. Each random object is place at a random position in the sky (i.e. ra, dec) Points with constant right ascension $\alpha = \text{const}$ draw great circles on the sky while points with constant declination $\delta = \text{const}$ draw small circles, with the exception of $\delta = 0$ i.e. the equator. If not properly accounted for, high declinations will be more likely positions due to the fact that the radii of the small circles decrease as $r \propto \cos\delta$. Consider the surface density $\eta = \frac{\mathrm{d}N}{\mathrm{d}A}$ on a spherical surface with the surface element $\mathrm{d}A$ given by

$$\mathrm{d}A = \cos\delta \mathrm{d}\delta\mathrm{d}\alpha. \tag{5.1}$$

   Introducing $u = \sin\delta \Leftrightarrow \mathrm{d}u = \cos\delta\mathrm{d}\delta$, the surface element is then given simply by $\mathrm{d}A = \mathrm{d}u\mathrm{d}\alpha$. This gives that as long as $u \in [-1, 1]$ and $\alpha \in [0, 2\pi]$ are drawn uniformly, the surface density will be constant across the sky, and the declination is then given by $\delta = \arcsin u$.

2. To each random object is assigned a flux randomly extracted from the data catalog. Allevato et al. (2011) showed that the flux selection produces the same results as if

[4] http://www.openmp.org/

extracting the simulated sources from a reference input $\log N - \log S$, i.e. number of sources as a function of flux. The random source is kept in the random sample if its flux is above the sensitivity map (figure 4.3, Cappelluti et al. 2012) value at that position. Otherwise the object is discarded and a new position is selected at random (step 1). The sensitivity map is given as a FITS image, where each pixel of the image has a limiting flux value assigned to it. Thus one needs to be able to map the astronomical coordinates to the pixel positions and vice versa. The exact formulation of this transformation is dependant on the projection used and the location of the field in the sky. However, a detailed look at the transformations involved is presented by Calabretta and Greisen (2002). A small caveat, while drawing the flux at random, is to take into account that the probability at drawing a given flux f_i could be proportional to the weight assigned to the flux i.e. $p(\text{flux} = f_i) \propto w_i$, where w_i is the weight of the flux f_i.

3. To each random object is assigned a redshift based on the redshift distribution of the data catalog. The redshift distribution of the data catalog contains selection effects; With increasing redshift, only more luminous objects are detected. In addition, the redshift distribution shows clustering in the form of sharp peaks over a small redshift bin Δz. In order to get rid of the clustering, the redshift distribution of the data catalog is smoothed and a redshift value is drawn from the smoothed redshift distribution. For example, Gilli et al. (2005, 2009); Allevato et al. (2011) have used a Gaussian smoothing with $\sigma_z = 0.3$, which is a compromise between not properly smoothing the spikes in the redshift distribution and excessive smoothing . In this work, a Kernel Density Estimation (KDE) method is used with a choice of a Gaussian kernel. Specifically, in the KDE method, each discrete redshift z_i is replaced by a Gaussian *kernel*:

$$p_i(\text{redshift} = z) = \frac{1}{\sqrt{2\pi}\sigma_z} \exp\left(-\frac{(z - z_i)^2}{2\sigma_z^2}\right), \tag{5.2}$$

where $\sigma_z = 0.3$, and the probability density function of redshift is then the sum of the kernels:

$$p(\text{redshift} = z) \propto \sum_i p_i(z). \tag{5.3}$$

If the probability distribution p_i sums to w_i i.e. $\sum_k p_i(\text{redshift} = z_k) = w_i$, then the normalization of equation 5.3 is $\sum_i w_i$ and the probability distribution function is:

$$p(\text{redshift} = z) = \frac{\sum_i p_i}{\sum_i w_i}. \tag{5.4}$$

Drawing from the smoothed redshift distribution may be then done utilizing the inverse CDF method (see section §3.3).

Figure 5.1 shows the redshift distribution of $n_{\text{obj}} = 632$ XMM-COSMOS AGN with known spectroscopic redshifts and the scaled random sample with $n_r = 100 n_{\text{obj}}$. For each of the subsample described in section §5.5, the random catalog is created individually with $n_r \sim 100 n_{\text{obj}}$ and using a Gaussian smoothing length $\sigma_z = 0.3$.

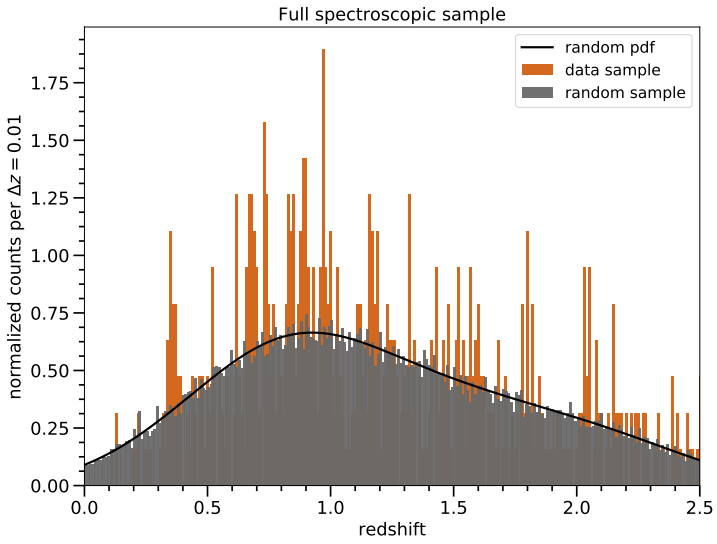

Figure 5.1: The redshift distribution of the full XMM-COSMOS AGN sample with known spectroscopic redshifts (red histogram). Shown also is the random sample (grey histogram) and the probabilitiy density function of the smoothed redshift distribution. The binsize is $\Delta z = 0.01$ for the histograms.

5.3 Error estimation

The uncertainty of the projected correlation function is estimated using the Bootstrap resampling technique. This is done efficiently by dividing the whole survey area into N_r cubic regions. Upon calculating the pairs, each pair is assigned into a region so that a total

of N regions have pairs assigned to them. Note that not every region necessarily contain pairs, which simply gives that $N \leq N_r$. Now, the resampling is performed by selecting N regions at random with resampling and assigning a weight to the region based on how many times it has been selected. The projected 2PCF $w_p(r_p)$ is then estimated again with the resampled regions. The resampling process is repeated k times, which gives the bootstrap distribution of $w_p(r_p)$. For each r_p, there exists k distinct values of w_p values. For these k values, the covariance matrix can then be readily estimated using equation 3.8.

In the following analysis the covariance matrix of $w_p(r_p)$ is estimated by dividing the survey area into $N_x = N_y = N_z = 10$, or $N_r = 1000$, cubic regions and using $k = 1000$ resampling trials. The errors of $w_p(r_p)$ are reported as the square root of the diagonal of the covariance matrix i.e. for the r_p bin i, the error is $\Delta w_p(r_p = r_p^i) = \sqrt{C_{i,i}}$.

5.4 $\pi_{\mathbf{max}}$ determination

The projected 2PCF (equation 2.35) is defined via the 2D spatial correlation function $\xi(r_p, \pi)$ (2D 2PCF), by integrating along the line-of-sight direction π. The separations perpendicular and parallel to the line-of-sight, r_p and π respectively, are given in comoving coordinates. Rather than integrating to infinity, the integration limit is a finite value $\pi_{\max}$. The appropriate value of $\pi_{\max}$ can be inferred by finding the value at which the AGN clustering signal levels off. Since the counting of pairs is computationally far more expensive than simple numerical integration, the $\pi_{\max}$ can be, in theory, determined with great accuracy. This is because the DD, DR and RR pairs only need to be counted once, after which the integration limit may be pruned relatively quickly over the range of interest.

The projected 2PCF for the $n_{\rm obj} = 632$ XMM-COSMOS AGN with known spectroscopic redshifts is shown in figure 5.2. From the figure, it may be inferred that the signal does not increase significantly after $\pi_{\max} \sim 40\ h^{-1}$Mpc, so the value $\pi_{\max} = 40\ h^{-1}$Mpc is adopted.

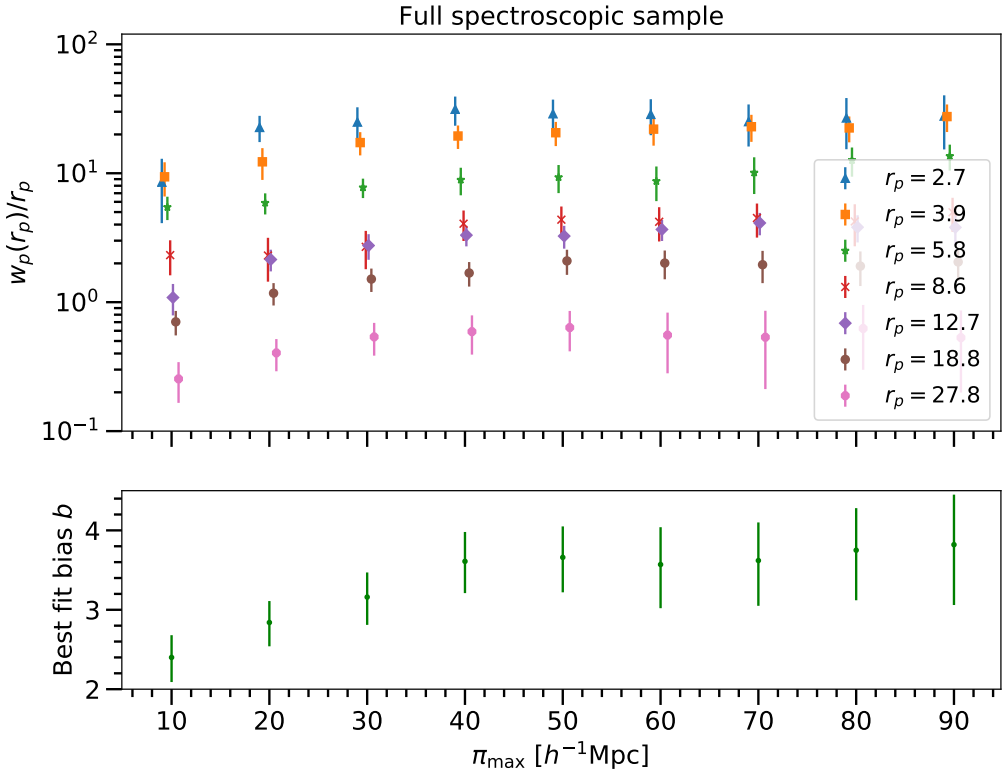

Figure 5.2: $\pi_{\max}$ determination for the full spectroscopic sample. The upper panel shows the projected correlation function (scaled by r_p for clarity) at different scales as a function of $\pi_{\max}$iwhile the lower panel shows the bias parameter b (equation 2.48) calculated using the χ^2 minimization technique (equation 3.4) for the given $\pi_{\max}$ value. The units of r_p are h^{-1}Mpc throughout. The correlation signal as well as the bias parameter b increase incrementally until $\pi_{\max} \sim 40\ h^{-1}$Mpc, above which no substantial increase is seen in the correlation signal, which means that the most of the correlated pairs are taken into account with $\pi_{\max} \sim 40\ h^{-1}$Mpc.

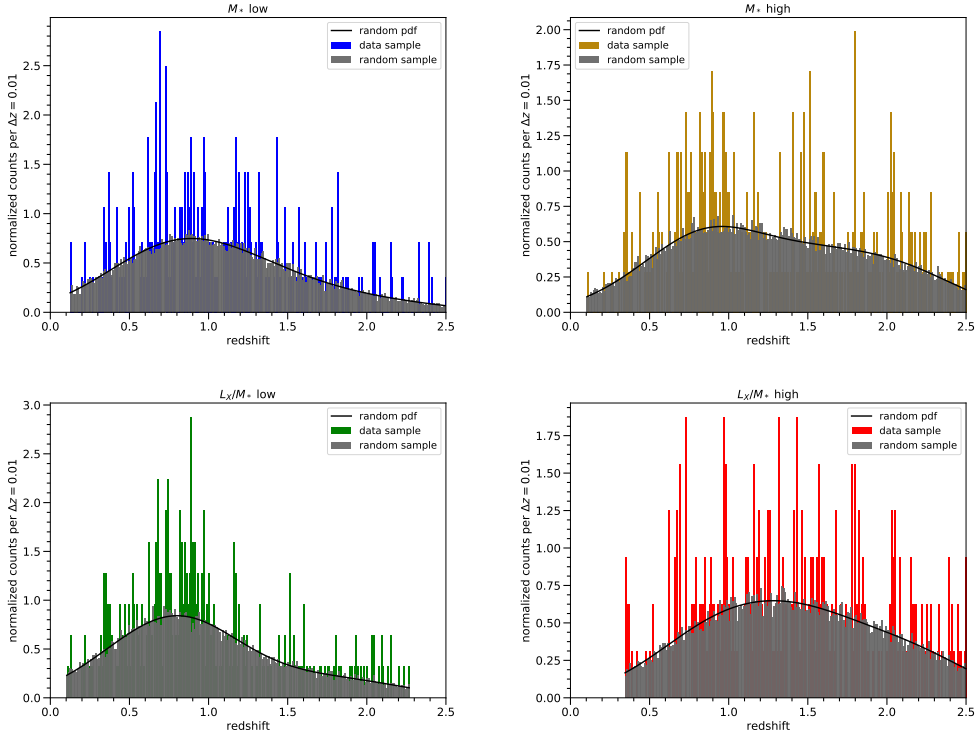

Figure 5.3: The redshift distributions of the subsamples and their respective random catalogs.

5.5 XMM-COSMOS AGN subsamples

The full sample of XMM-COSMOS AGN with known spectroscopic redshifts spans over a range of $0.1 < z < 2.5$ with $\overline{z} = 1.19$ and consists of $n_{\mathrm{obj}} = 632$. Figures 5.4 and 5.5 show the distribution of the host galaxy stellar mass (in units of $\mathrm{M}_\odot$) and the ratio between the X-ray luminosity and the host stellar mass i.e. a quantity proportional to the Eddington ratio in units of erg s^{-1} $\mathrm{M}_\odot^{-1}$. The mean values are $\log \overline{M_*} = 10.96$ and $\log \overline{L_X/M_*} = 33.56$ given in their described units.

The clustering of most of the objects with spectroscopic redshifts has already been well-studied by e.g. Gilli et al. (2009) and Allevato et al. (2011), which means that the spectroscopic sample may be used as a sanity check to verify the analysis process. Further, the sample is divided into subsamples, so that the number of objects in each subsample is roughly equal, based on the host galaxy stellar mass or the Eddington ratio. In table 5.1, the properties of each subsample are presented.

In detail, the low (high) M_* sample consists of 281 (351) AGN with $\log M_* \leq$

sample	criterion	$n_{\rm obj}$	$\overline{z}$	$\log \overline{M}_*$	$\log \overline{L_X/M_*}$
specz	full spectroscopic sample	632	1.19	10.96	33.56
M_* low	$\log M_* \leq 10.72$	281	1.05	10.40	33.80
M_* high	$\log M_* > 10.72$	351	1.30	11.16	33.18
L_X/M_* low	$\log L_X/M_* \leq 33.017$	312	0.96	11.08	32.59
L_X/M_* high	$\log L_X/M_* > 33.017$	320	1.41	10.81	33.83

Table 5.1: Subsample summary. M_* is in units of $\mathrm{M}_\odot$ and L_X/M_* in units of erg $\mathrm{s}^{-1}/\mathrm{M}_\odot$.

10.72 ($\log M_* > 10.72$). The low (high) Eddington ratio objects are 312 (320) and have $\log L_X/M_* \leq 33.017$ ($\log L_X/M_* > 33.017$). Note that a ratio of $\log L_X/M_* \sim 10^{34}$ corresponds roughly to an object accerting at the Eddington limit $\lambda_{\rm Edd} \sim 1$ (Bongiorno et al., 2012). The M_* and Eddington ratio subsamples are presented in figure 5.4 and figure 5.5, respectively.

5.6 Clustering measurements

The 2D 2PCF $\xi(r_p, \pi)$ has been measured in the range $r_p = 1 - 100$ h^{-1}Mpc in 10 logarithmic bins and $\pi = 0 - 100$ h^{-1}Mpc in 20 linear bins for the full sample and for the different subsamples. Then $\xi(r_p, \pi)$ is integrated up to different $\pi_{\rm max}$ values (in bins of $\Delta\pi_{\rm max} = 10$ h^{-1}Mpc) and the optimal value, where the projected 2PCF $w_p(r_p)$ converges has been found for each sample. Figure 5.6 and 5.7 show $w_p(r_p)$ in the range $r_p = 2 - 50$ h^{-1}Mpc for the diverse XMM-COSMOS AGN subsets. The 1σ errors are the square root of the diagonal components of the covariance matrix, which quantifies the correlation among different bins. The continuous line in the figures represents the AGN two-halo term (equation 2.48), where the DM 2PCF (dashed line) is estimated at the mean redshift of each sample by using a dark matter power spectrum shape parameter $\Gamma = 0.2$. Following equation 2.48, the best-fit bias is derived by using a χ^2 minimization technique with one free parameter (section 3.1). Here, the $\chi^2_{\rm min}$ value corresponds to the best-fit bias and the 1σ errors for the bias are found via $\Delta\chi^2 = \chi^2_{\rm min} + 1.0$, for one free parameter. The

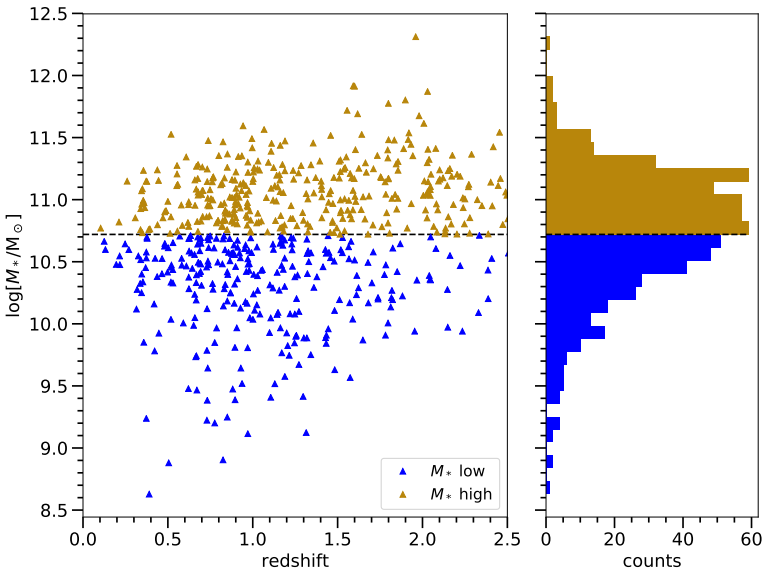

Figure 5.4: The M_* distribution as a function of redshift. The dashed line is the dividing value between the low subsample and the high subsample. The selection effect is clearly visible on the data, as at higher redshifts, only the more massive galaxies are sampled. The width of the bins in the histogram is $\Delta \log M_* = 0.1$.

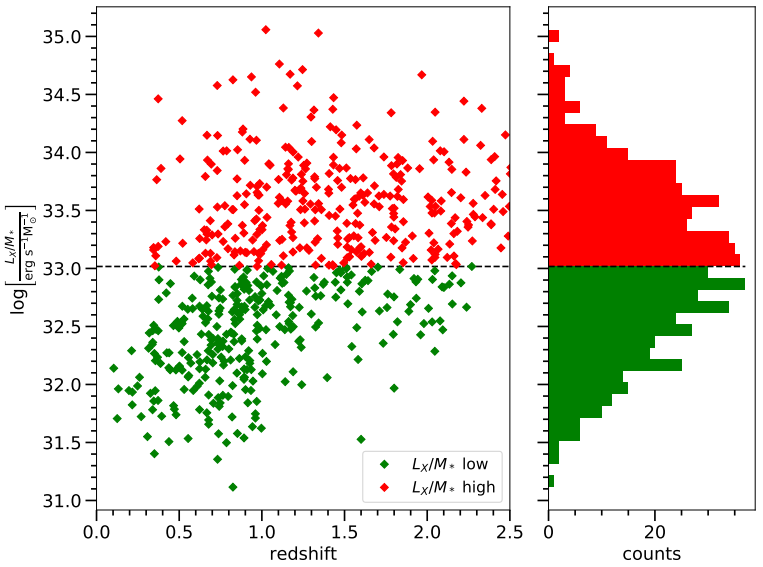

Figure 5.5: The L_X/M_* distribution as a function of redshift. The dashed line divides the low and the high subsample. As with the M_* distribution, at higher redshifts only the AGN with more intense accretion rates are sampled. The width of the bins in the histogram is $\Delta \log L_X/M_* = 0.1$.

linear bias of the halo model only applies on large scales, so the range for the fit used is $r_p = 2 - 50\ h^{-1}$Mpc.

Following section §2.3, the bias is then related to a typical mass of the hosting halo by using the ellipsoidal collapse model by Sheth et al. (2001) assuming $\delta_c = 1.686$, which is the critical density threshold for collapse. The peak height parameter $\nu \equiv \delta_c/\sigma$ parameter, is estimated using the prescription provided by van den Bosch (2002) (see appendix A) assuming a power spectrum shape parameter $\Gamma = 0.2$. This allows one to calculate the bias evolution of a constant halo mass i.e. $b(z, M_{\rm halo} = \text{const})$, and it is possible to invert the relation in order to reconcile the halo mass for a given bias i.e. $M_{\rm halo}(b, z = \overline{z})$, where b is the bias from the 2-halo model and $\overline{z}$ is the mean redshift of the subsample. However, this approach only reveals the *characteristic* halo mass for the given AGN population and the mass of the hosting halo for any single given AGN may indeed differ from the characteristic halo mass and the mass of the hosting halo holds in a statistical sense for the sample.

For the spectroscopic sample with $\overline{z} = 1.19$ and $n_{\rm obj} = 632$ the χ^2 minimization yields a best-fit bias parameter of $b = 3.62 \pm 0.13$ (see figure 5.6), which corresponds to a typical mass of the hosting halo of $\log M_{\rm halo}/h^{-1}{\rm M}_\odot = 13.52 \pm 0.05$. On scales $r_p \sim 10 - 30$ h^{-1}Mpc the clustering signal is not taken well into account by the model and likely thus $\chi^2_{\rm min}/\nu \geq 1$. The excess in the signal on these scales might be due to a large structure present in the data, or alternatively (or simultaneously), the high-z part of the sample dominates the clustering signal on the large scale, with a larger respective bias.

The low M_* subsample with $\overline{z} = 1.05$ and $n_{\rm obj} = 281$ yields a best-fit bias of $b = 3.61^{+0.37}_{-0.40}$ which corresponds to a characteristic halo mass of $M_{\rm halo} = 13.52^{+0.12}_{-0.16}$. A similar typical hosting halo mass is found for the high M_* subsample with $\overline{z} = 1.30$, $n_{\rm obj} = 351$ and $b = 3.53^{+0.58}_{-0.70}$, for which the corresponding halo mass is $M_{\rm halo} = 13.61^{+0.19}_{-0.30}$.

The L_X/M_*, i.e. the eddington ratio, subsamples L_X/M_* low with $\overline{z} = 0.96$, $n_{\rm obj} = 312$ and L_X/M_* high with $\overline{z} = 1.41$, $n_{\rm obj} = 320$, have best-fit biases $b = 3.09^{+0.57}_{-0.69}$ (low subsample) and $b = 3.51^{+0.84}_{-1.11}$ (high subsample). The corresponding halo masses are $M_{\rm halo} = 13.52^{+0.22}_{-0.37}$ and $M_{\rm halo} = 13.29^{+0.28}_{-0.58}$ for the low subsample and the high subsample, respectively. This would *tentatively* suggest that higher halo masses are associated with

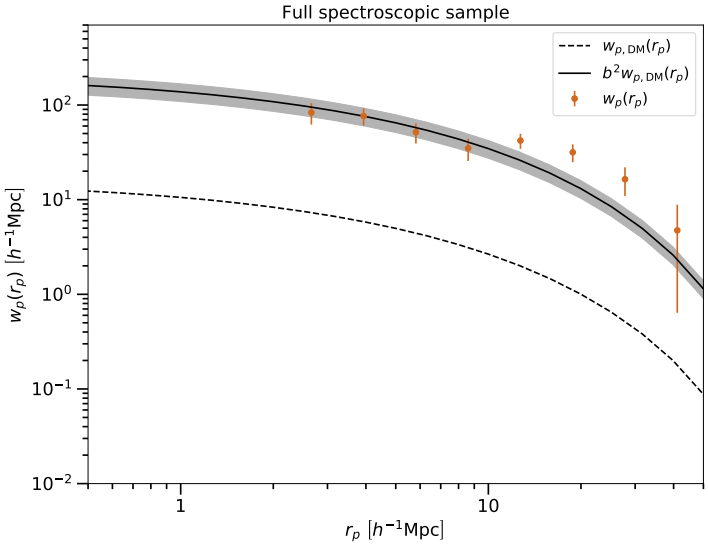

Figure 5.6: The projected correlation function for the $n_{\mathrm{obj}} = 632$ XMM-COSMOS AGN with known spectroscopic redshifts. The errors on the data points are the square root of the diagonal elements of the covariance matrix using the bootstrap resampling method. The dashed line is the projected 2PCF for the underlying DM distribution at $z = 1.2$, and the solid black line indicates the best-fit bias found via χ^2 fitting and the shaded region corresponds to the 1σ errors for the best-fit bias.

lower values of the Eddington ratio. Such conclusions should be made with great care as there are selection effects at play in the data. As is evident from the mean redshifts of the subsamples and figure 5.5, the L_X/M_* high subsample consists of objects at higher redshifts, due to objects accreting at higher rates being also more luminous.

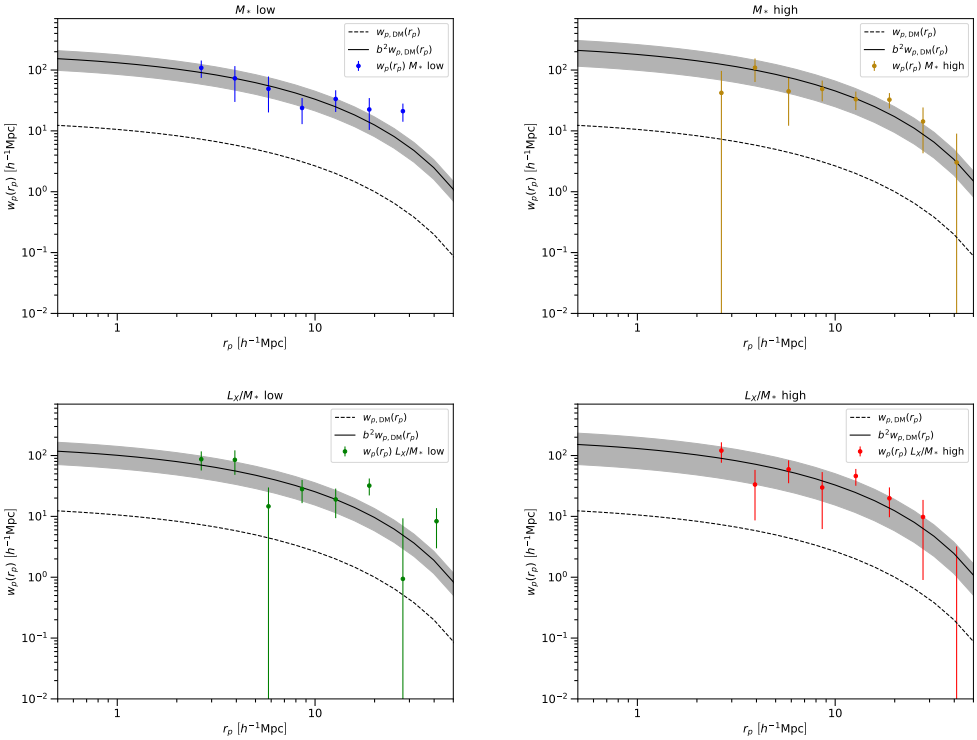

Figure 5.7: The projected 2PCF $w_p(r_p)$ for the respective subsamples. The continous lines represent the AGN two halo term (equation 2.48), where the DM 2PCF (dashed lines) is estimated at the mean redshift of each sample. The errors for the data points are the square root of the diagonal elements of the covariance matrix estimated using the bootstrap resampling method.

sample	$\pi_{\max}$	$\overline{z}$	$\chi^2_{\min}/\nu$	b (eq. 2.48)	$\log M_{\rm halo}$
specz	40	1.19	3.03	$3.61^{+0.37}_{-0.40}$	$13.52^{+0.12}_{-0.16}$
M_* low	40	1.05	4.04	$3.53^{+0.58}_{-0.70}$	$13.61^{+0.19}_{-0.30}$
M_* high	40	1.30	1.32	$4.13^{+0.85}_{-1.07}$	$13.60^{+0.22}_{-0.40}$
L_X/M_* low	40	0.96	2.09	$3.09^{+0.57}_{-0.69}$	$13.52^{+0.22}_{-0.37}$
L_X/M_* high	40	1.41	1.62	$3.51^{+0.84}_{-1.11}$	$13.29^{+0.28}_{-0.58}$

Table 5.2: Results summary. $\pi_{\max}$ is given in units of h^{-1}Mpc and $M_{\rm halo}$ in units of $h^{-1}\mathrm{M}_\odot$. The reported errors are 1σ errors.

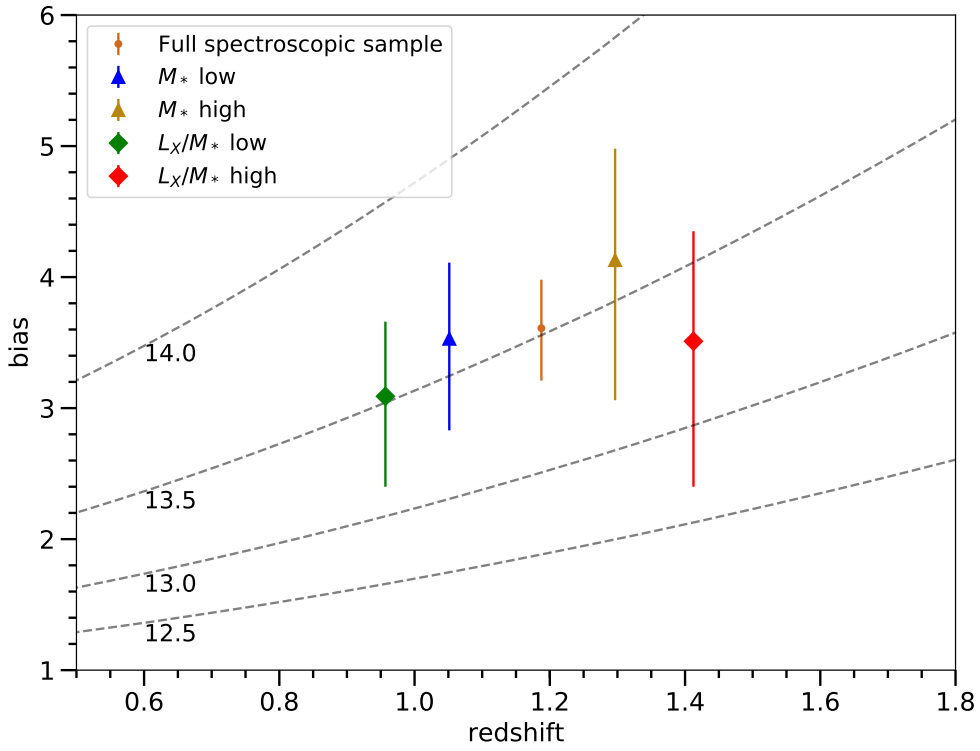

Figure 5.8: Bias factors versus redshift for the full XMM-COSMOS AGN sample with known spectroscopic redshifts, and the M_* (triangles) and L_X/M_* (diamonds) subsamples. The dashed lines correspond to constant halo mass thresholds, i.e. $b(z, M_{\rm halo} = \text{const})$, from Sheth et al. (2001) and van den Bosch (2002) in logarithmic units of $h^{-1}\mathrm{M}_\odot$.

6. Discussion and conclusions

The clustering properties of $n_{\rm obj} = 632$ XMM-COSMOS AGN have been studied, with known spectroscopic redshifts in the range $0.1 < z < 2.5$, as a function of the host galaxy properties, such as the stellar mass and the Eddington ratio. The bias factor has been estimated using the AGN projected 2PCF at large scales relative to the underlying DM distribtuion and then converted into a characteristic mass of the hosting haloes following the ellipsoidal collapse model of Sheth et al. (2001). It was found that XMM-COSMOS AGN at $\overline{z} = 1.2$ reside in massive haloes $\log M_{\rm halo}/h^{-1}{\rm M}_{\odot} = 13.52^{+0.12}_{-0.16}$, i.e. typical in dense environments, such as galaxy groups.

The halo masses have been estimated for similar samples of X-ray selected XMM-COSMOS AGN by Gilli et al. (2009) and by Allevato et al. (2011). Gilli et al. (2009) associated the AGN with haloes from the Millenium Simulation and found that they are likely to reside in haloes with masses $\gtrsim 10^{12}$ M$_{\odot}$ while Allevato et al. (2011) found, by using the halo model, that the AGN sample is characterized by a halo mass of $\log M_{\rm halo}/h^{-1}{\rm M}_{\odot} = 13.23 \pm 0.06$. The characteristic halo mass derived for the sample used in this work differs slightly from that of Allevato et al. (2011), it is consistent with the overall picture of X-ray selected AGN residing in massive DM haloes with $12.5 < \log M_{\rm halo}/h^{-1}{\rm M}_{\odot} < 13.5$ (e.g. Cappelluti et al. 2012). Following Allevato et al. (2011), the DM haloes are more massive than predicted by theoretical models that assume quasar phase triggered by major merger, which suggest instead characteristic masses of $\log M/h^{-1}{\rm M}_{\odot} \sim 12$ in this redshift range.

XMM-COSMOS AGN with low and high M_* were found to reside in DM haloes with similar masses. However, the redshift evolution of the bias is not taken into account as the low and high M_* samples have different mean redshifts, $\overline{z} = 1.05$ versus $\overline{z} = 1.30$.

On the contrary, the low and high L_X/M_* subsamples show a similar bias, but taking into account the different average redshifts $z = 0.96$ and $z = 1.41$ for the low and high subsample, respectively, there is a a slight difference in the mass of the hosting haloes with the low L_X/M_* subsample featuring a slightly higher typical halo mass. A possible explanation for this could be that the environment of the low L_X/M_* reduces the gaseous content available for accretion. Lower amounts of gas would then result in lower accretion rates, i.e. low L_X/M_*. This presents an interesting future prospect for a study, where the selection effects in the sample could be taken better into account by studying the accretion rates in smaller redshift bins.

The analysis performed in chapter §5 serves as the beginning point for a future study. The catalog in this analysis utilized only the AGN with spectroscopic redshifts with $0.1 < z < 2.5$. However, with the inclusion of readily available photometric redshifts, the number of objects may be effectively doubled, which, in theory should improve the quality of the clustering signal, especially in the high-z Universe, where spectra may be scarce compared to photometry. Preliminary results obtained by measuring the projected 2PCF using both spectroscopic redshifts and any additional photometric redshifts available are very promising, but also suggest that the effects of introducing photometric probability distribution functions are rather complex. The effect of introducing photometric redshifts in clustering measurements have been studied only in a few recent papers (e.g. Georgakakis et al. 2014, Allevato et al. 2016) and are likely important for clustering measurements in the future using large X-ray AGN surveys, where the fraction of AGN with secure spectroscopic redshifts might be small.

Bibliography

Allevato, V., Civano, F., Finoguenov, A., et al. The Chandra COSMOS Legacy Survey: Clustering of X-Ray-selected AGNs at $2.9 \leq z \leq 5.5$ Using Photometric Redshift Probability Distribution Functions. *ApJ*, 832:70, 2016. doi:10.3847/0004-637X/832/1/70.

Allevato, V., Finoguenov, A., Cappelluti, N., et al. The XMM-Newton Wide Field Survey in the COSMOS Field: Redshift Evolution of AGN Bias and Subdominant Role of Mergers in Triggering Moderate-luminosity AGNs at Redshifts up to 2.2. *ApJ*, 736:99, 2011. doi:10.1088/0004-637X/736/2/99.

Alonso, D. CUTE solutions for two-point correlation functions from large cosmological datasets. *ArXiv e-prints*, 2012.

Antonucci, R. Unified models for active galactic nuclei and quasars. *ARAA*, 31:473–521, 1993. doi:10.1146/annurev.aa.31.090193.002353.

Avni, Y. Energy spectra of X-ray clusters of galaxies. *ApJ*, 210:642–646, 1976. doi:10.1086/154870.

Bongiorno, A., Merloni, A., Brusa, M., et al. Accreting supermassive black holes in the COSMOS field and the connection to their host galaxies. *MNRAS*, 427:3103–3133, 2012. doi:10.1111/j.1365-2966.2012.22089.x.

Brandt, W. N. and Hasinger, G. Deep Extragalactic X-Ray Surveys. *ARAA*, 43:827–859, 2005. doi:10.1146/annurev.astro.43.051804.102213.

Brusa, M., Civano, F., Comastri, A., et al. The XMM-Newton Wide-field Survey in the Cosmos Field (XMM-COSMOS): Demography and Multiwavelength Properties of Obscured and Unobscured Luminous Active Galactic Nuclei. *ApJ*, 716:348–369, 2010. doi:10.1088/0004-637X/716/1/348.

Brusa, M., Zamorani, G., Comastri, A., et al. The XMM-Newton Wide-Field Survey in the COSMOS Field. III. Optical Identification and Multiwavelength Properties of a Large Sample of X-Ray-Selected Sources. *ApJS*, 172:353–367, 2007. doi:10.1086/516575.

Calabretta, M. R. and Greisen, E. W. Representations of celestial coordinates in FITS. *AAP*, 395:1077–1122, 2002. doi:10.1051/0004-6361:20021327.

Cappelluti, N., Allevato, V., and Finoguenov, A. Clustering of X-Ray-Selected AGN. *Advances in Astronomy*, 2012:853701, 2012. doi:10.1155/2012/853701.

Cappelluti, N., Brusa, M., Hasinger, G., et al. The XMM-Newton wide-field survey in the COSMOS field. The point-like X-ray source catalogue. *AAP*, 497:635–648, 2009. doi:10.1051/0004-6361/200810794.

Civano, F., Marchesi, S., Comastri, A., et al. The Chandra Cosmos Legacy Survey: Overview and Point Source Catalog. *ApJ*, 819:62, 2016. doi:10.3847/0004-637X/819/1/62.

Coil, A. L. *The Large-Scale Structure of the Universe*, page 387. 2013. doi:10.1007/978-94-007-5609-0_8.

Cooray, A. and Sheth, R. Halo models of large scale structure. *Physics Reports*, 372:1–129, 2002. doi:10.1016/S0370-1573(02)00276-4.

Davis, M. and Peebles, P. J. E. A survey of galaxy redshifts. V - The two-point position and velocity correlations. *ApJ*, 267:465–482, 1983. doi:10.1086/160884.

Elvis, M., Civano, F., Vignali, C., et al. The Chandra COSMOS Survey. I. Overview and Point Source Catalog. *ApJS*, 184:158–171, 2009. doi:10.1088/0067-0049/184/1/158.

Fanaroff, B. L. and Riley, J. M. The morphology of extragalactic radio sources of high and low luminosity. *MNRAS*, 167:31P–36P, 1974. doi:10.1093/mnras/167.1.31P.

Georgakakis, A., Mountrichas, G., Salvato, M., et al. Large-scale clustering measurements with photometric redshifts: comparing the dark matter haloes of X-ray AGN, star-forming and passive galaxies at z $\sim$ 1. *MNRAS*, 443:3327–3340, 2014. doi:10.1093/mnras/stu1326.

Gilli, R., Daddi, E., Zamorani, G., et al. The spatial clustering of X-ray selected AGN and galaxies in the Chandra Deep Field South and North. *AAP*, 430:811–825, 2005. doi:10.1051/0004-6361:20041375.

Gilli, R., Zamorani, G., Miyaji, T., et al. The spatial clustering of X-ray selected AGN in the XMM-COSMOS field. *AAP*, 494:33–48, 2009. doi:10.1051/0004-6361:200810821.

Gültekin, K., Richstone, D. O., Gebhardt, K., et al. The M-σ and M-L Relations in Galactic Bulges, and Determinations of Their Intrinsic Scatter. *ApJ*, 698:198–221, 2009. doi:10.1088/0004-637X/698/1/198.

Gunn, J. E. and Gott, J. R., III. On the Infall of Matter Into Clusters of Galaxies and Some Effects on Their Evolution. *ApJ*, 176:1, 1972. doi:10.1086/151605.

Hasinger, G., Cappelluti, N., Brunner, H., et al. The XMM-Newton Wide-Field Survey in the COSMOS Field. I. Survey Description. *ApJS*, 172:29–37, 2007. doi:10.1086/516576.

Hickox, R. C., Jones, C., Forman, W. R., et al. Host Galaxies, Clustering, Eddington Ratios, and Evolution of Radio, X-Ray, and Infrared-Selected AGNs. *ApJ*, 696:891–919, 2009. doi:10.1088/0004-637X/696/1/891.

Jenkins, A., Frenk, C. S., White, S. D. M., et al. The mass function of dark matter haloes. *MNRAS*, 321:372–384, 2001. doi:10.1046/j.1365-8711.2001.04029.x.

Kaiser, N. Clustering in real space and in redshift space. *MNRAS*, 227:1–21, 1987. doi:10.1093/mnras/227.1.1.

Laureijs, R., Amiaux, J., Arduini, S., et al. Euclid Definition Study Report. *ArXiv e-prints*, 2011.

Longair, M. S. *High Energy Astrophysics*. 2011.

Marulli, F., Veropalumbo, A., and Moresco, M. CosmoBolognaLib: C++ libraries for cosmological calculations. *Astronomy and Computing*, 14:35–42, 2016. doi:10.1016/j.ascom.2016.01.005.

Miyaji, T., Zamorani, G., Cappelluti, N., et al. The XMM-Newton Wide-Field Survey in the COSMOS Field. V. Angular Clustering of the X-Ray Point Sources. *ApJS*, 172:396–405, 2007. doi:10.1086/516579.

Mo, H., van den Bosch, F. C., and White, S. *Galaxy Formation and Evolution*. Cambridge University Press, 2010.

Moore, A. W., Connolly, A. J., Genovese, C., et al. Fast Algorithms and Efficient Statistics: N-Point Correlation Functions. In A. J. Banday, S. Zaroubi, and M. Bartelmann, editors, *Mining the Sky*, page 71. 2001. doi:10.1007/10849171_5.

Mortlock, D. J., Warren, S. J., Venemans, B. P., et al. A luminous quasar at a redshift of z = 7.085. *NATURE*, 474:616–619, 2011. doi:10.1038/nature10159.

NASA and The Hubble Heritage Team (STScI/AURA). The M87 Jet. 2000. [Online; accessed August 21, 2017].

Norberg, P., Baugh, C. M., Gaztañaga, E., and Croton, D. J. Statistical analysis of galaxy surveys - I. Robust error estimation for two-point clustering statistics. *MNRAS*, 396:19–38, 2009. doi:10.1111/j.1365-2966.2009.14389.x.

NRAO/AUI. Quasar 3C 175. 1996. [Online; accessed August 21, 2017].

Peacock, J. A., Cole, S., Norberg, P., et al. A measurement of the cosmological mass density from clustering in the 2dF Galaxy Redshift Survey. *NATURE*, 410:169–173, 2001.

Peebles, P. J. E. *The large-scale structure of the universe.* 1980.

Perlmutter, S., Aldering, G., Goldhaber, G., et al. Measurements of Ω and Λ from 42 High-Redshift Supernovae. *ApJ*, 517:565–586, 1999. doi:10.1086/307221.

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. Planck 2015 results. XIII. Cosmological parameters. *AAP*, 594:A13, 2016. doi:10.1051/0004-6361/201525830.

Press, W. H. and Schechter, P. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. *ApJ*, 187:425–438, 1974. doi:10.1086/152650.

Riess, A. G., Filippenko, A. V., Challis, P., et al. Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *AJ*, 116:1009–1038, 1998. doi:10.1086/300499.

Roos, M. *Introduction to Cosmology, Third Edition.* 2003.

Salvato, M., Hasinger, G., Ilbert, O., et al. Photometric Redshift and Classification for the XMM-COSMOS Sources. *ApJ*, 690:1250–1263, 2009. doi:10.1088/0004-637X/690/2/1250.

Salvato, M., Ilbert, O., Hasinger, G., et al. Dissecting Photometric Redshift for Active Galactic Nucleus Using XMM- and Chandra-COSMOS Samples. *ApJ*, 742:61, 2011. doi:10.1088/0004-637X/742/2/61.

Schmidt, M. 3C 273 : A Star-Like Object with Large Red-Shift. *NATURE*, 197:1040, 1963. doi:10.1038/1971040a0.

Scoville, N., Aussel, H., Brusa, M., et al. The Cosmic Evolution Survey (COSMOS): Overview. *ApJS*, 172:1–8, 2007. doi:10.1086/516585.

Sheth, R. K., Mo, H. J., and Tormen, G. Ellipsoidal collapse and an improved model for the number and spatial distribution of dark matter haloes. *MNRAS*, 323:1–12, 2001. doi:10.1046/j.1365-8711.2001.04006.x.

Sheth, R. K. and Tormen, G. Large-scale bias and the peak background split. *MNRAS*, 308:119–126, 1999. doi:10.1046/j.1365-8711.1999.02692.x.

Springel, V., White, S. D. M., Jenkins, A., et al. Simulations of the formation, evolution and clustering of galaxies and quasars. *NATURE*, 435:629–636, 2005. doi:10.1038/nature03597.

Tinker, J. L., Weinberg, D. H., Zheng, Z., and Zehavi, I. On the Mass-to-Light Ratio of Large-Scale Structure. *ApJ*, 631:41–58, 2005. doi:10.1086/432084.

Urry, C. M. and Padovani, P. Unified Schemes for Radio-Loud Active Galactic Nuclei. *PASP*, 107:803, 1995. doi:10.1086/133630.

van den Bosch, F. C. The universal mass accretion history of cold dark matter haloes. *MNRAS*, 331:98–110, 2002. doi:10.1046/j.1365-8711.2002.05171.x.

Wall, J. V. and Jenkins, C. R. *Practical Statistics for Astronomers.* Second edition, 2012.